The Proteomic Landscape of Human Disease: Construction and Evaluation of Networks Associated to Complex Traits

Dr. Mark Joseph Daly Elizabeth Jeffries Rossin

The proteomic landscape of human disease: construction and evaluation of networks associated to complex traits

Abstract

Genetic mapping of complex traits has been successful over the last decade, with over 2,000 regions in the genome associated to disease. Yet, the translation of these findings into a better understanding of disease biology is not straightforward. The true promise of human genetics lies in its ability to explain disease etiology, and the need to translate genetic findings into a better understanding of biological processes is of great relevance to the community. We hypothesized that integrating genetics and protein-protein interaction (PPI) networks would shed light on the relationship among genes associated to complex traits, ultimately to help guide understanding of disease biology.

First, we discuss the design, testing and implementation of a novel *in silico* approach ("DAPPLE") to rigorously ask whether loci associated to complex traits code for proteins that form significantly connected networks. Using a high-confidence set of publically available physical interactions, we show that loci associated to autoimmune diseases code for proteins that assemble into significantly connected networks and that these networks are predictive of new genetic variants associated to the phenotypes in question.

Next, we study variation in the electrocardiographic QT-interval, a heritable phenotype that when prolonged is a risk factor for cardiac arrhythmia and sudden cardiac death. We show that a large proportion of QT-associated loci encode proteins that are members of complexes identified by immunoprecipitations in mouse cardiac tissue of

proteins known to be causal of Mendelian long-QT syndrome. For several of the identified proteins, we show they affect cardiac ion channel currents in model organisms. Using replication genotyping in 17,500 individuals, we use the complexes to identify genome-wide significant loci that would have otherwise been missed.

Finally, we consider whether PPIs can be used to interpret rare and *de novo* variation discovered through recent technological advances in exome-sequencing. We report a highly connected network underlying *de novo* variants discovered in an autism trio exome-sequencing effort, and we design, test and implement a novel statistical framework ("DAPPLE/SEQ") to analyze rare inherited variants in the context of PPIs in a way that significantly boosts power to detect association.

Table of Contents

List of Figures

List of Tables

Thank you

There are many people who have helped me get to where I am today. I would like to thank my advisor Dr. Mark J Daly for teaching me human genetics, for his unwavering support of my progress and for his genuine enthusiasm for the field of human genetics and his purity as a scientist; Dr. Chris Cotsapas, who has taken me under his wing from day one and been a constant source of career and scientific advice; Dr. Kasper Lage for teaching me proteomics and for being a huge supporter of my work; Dr. Soumya Raychaudhuri for his continued support of my career as a physician-scientist; Dr. Benjamin Neale for his dedication to teaching me and others genetics and statistics; Dr. David Altshuler for his professional, scientific and career advice; Dr. Stephan Ripke for his daily collaboration and teaching; Dr. Ramnik Xaiver for his expert guidance on all projects; my Dissertation Advisory Committee (Dr. Joel Hirschhorn, Dr. Vamsi Mootha and Dr. Paul de Bakker); my defense committee (Dr. Joel Hirschhorn, Dr. Barbara Stranger, Dr. Aviv Regev and Dr. Peter Kraft); my family for their continued support and enthusiasm; my partner Vijay Patel for his unwavering encouragement and commitment and all my friends from near and far for being such a huge part of my life.

1 Introduction

1.1 Genetic Mapping in Human Disease

Studying DNA variation of people affected by heritable disorders is one of the most promising approaches to identifying the cellular causes of many human diseases. Especially for diseases whose etiology is largely unknown, finding potentially causal genes can help reveal proteins and pathways to target with therapeutics. The search for such genetic variation has been particularly successful in rare and highly penetrant genetic disorders that are caused by severe mutations in DNA: classic examples of such diseases include hemochromatosis, cystic fibrosis and phenylketonuria[1]. For these diseases, single nucleotide changes in particular genes lead to deficient or altered proteins that then result in a cascade of physiological outcomes, ultimately culminating in the medical sequelae that define the disease. Understanding the cellular processes that have gone awry is relevant to patients' medical treatment – for example, understanding that phenylketonuria is caused by a lesion the gene encoding phenylalanine hydroxylase directly translates to phenylalanine-lowering recommendations that drastically minimize damage; likewise, knowing that cystic fibrosis is caused by a primary defect in chloride transport has helped steer recent design of treatment toward drugs that will specifically focus on chloride ions. Therefore, we study human genetics to understand disease biology with the goal ultimately being novel therapeutic design.

One of the goals of the last decade has been to apply these concepts to more common and genetically complex diseases. A complex trait is defined as one that is influenced by many low-pentrant DNA variants as well as the environment such that the familial clustering of the trait does not follow a clear and predictable inheritance pattern.

For most complex diseases, we do not understand the bulk of the underlying pathophysiology, though many of these traits are clearly heritable. Since the early 1900s, medical doctors and scientists have compared monozygotic twins to dizygotic twins to estimate the proportion of phenotypic variance observed in such complex traits that is due to genetics. For many continuous traits – height, body mass index, blood lipid levels – as well as dichotomous traits – autoimmune disease, Type 2 diabetes and psychiatric disease – twin studies have consistently revealed a striking degree of heritability (typically ranging from 30-90%). This observation is what gives the field of human genetics (and in particular complex trait genetics) hope: identifying genetic risk variants is a reliable and unbiased means of gaining insight into the relevant genes and biological processes.

Identifying genes for Mendelian and complex traits alike requires genetic mapping, i.e. the identification and localization of genes that underlie phenotypes [2]. Genetic mapping is accomplished by correlating DNA variation that is nearby to the disease-causing mutation – single nucleotide polymorphisms (SNPs), microsatellite and minisatellites, copy number variation, etc – with phenotype. The fundamental concept of genetic mapping involves two observations: DNA gets passed on in blocks and only rarely undergoes recombination, and this recombination occurs at "hotspots", such that there are places in the genome that are much more likely to endure a recombination event than others. The first observation is what led to linkage mapping: rare and highly deleterious alleles could be tracked in families by correlating genomic markers (DNA variation) with phenotype, since variants close to the disease-causing mutation will rarely recombine and those that are far will do so more often – thus allowing for positional mapping of the disease region. The second observation means that correlation between

nearby variants bounded by recombination hotspots will persist from ancestors many generations ago since recombination between them will be rare. This phenomenon is known as linkage disequilibrium (LD), and LD is harnessed in genome-wide association studies (GWAS), discussed below. Both linkage and GWAS require being able to systematically assay a set of variants throughout the genome.

The earliest attempts at linkage mapping were carried out by Surtevant in the early 1900s in Drosophila, when he realized that he could map the linear order of genes by tracking patterns of correlations between genotype and phenotype in fly crosses, with the assumption that meiotic cross-overs would lead to association only between markers physically near to the phenotype-causing mutation (also known as "positional cloning")[3]. Though successful in model organisms where predetermined crosses could be carried out, such an approach was met with significant difficulties in humans for most of the 20th century due to small pedigrees and lack of systematic markers throughout the genome. Linkage in families became feasible around 1980 when Botstein and colleagues proposed the idea of using restriction fragment length polymorphisms (a type of variant that disrupts a restriction enzyme cut site and is therefore easy to assay) throughout the genome to systematically map human genes associated with disease[4]. This breakthrough in methodology led to the mapping of the Huntington's gene on chromosome 4 in 1983[5] followed by the systemic documentation of dense genome-wide polymorphic sites and the subsequent mapping of now over 2,000 Mendelian diseases[6]. Disease and study characteristics that aided in successful linkage mapping included highly penetrant causal genetic variants (as is the case in the aforementioned diseases like cystic fibrosis), causal variants that are rare in the population, relatively

little environmental influence on the phenotype, large families and minimal locus heterogeneity. Linkage analysis works very well with Mendelian phenotypes for these reasons.

As the number of Mendelian phenotypes successfully mapped through linkage rose, however, it quickly became clear that this approach was woefully underperforming in more common complex phenotypes that clearly exhibited a heritable component. As discussed, these traits are unlike Mendelian phenotypes in that they are highly polygenic (many risk variants throughout the genome), influenced by environment, and contributed to by genetic variants of individually very low effect on phenotype. These factors are what make such traits ill-suited for linkage analysis.

Association analysis in large samples of individuals – comparison of allele frequencies in cases and controls rather than tracking genotypes and phenotypes in a family – offered a path forward because it was much better powered to detect such associations of weak effects due to the fact that it does not require large families. Association analysis differs from linkage in that it simply looks for differences in allele frequencies between cases and controls, rather than positionally mapping a linked region. It turns out that association studies for common variation ended up harnessing the same concept as linkage (i.e., variants close to the causal mutation will tend to associated with phenotype), but are fundamentally distinct in that they do not consider meiotic crosses in families. Rather, they use variants that tag regions of LD throughout the genome and that have persisted from ancestors many generations ago and undergone relatively little recombination. The first attempts at this were based on candidate gene studies where investigators compared variants between cases and controls within a single gene of

interest, such as the mapping of the human leukocyte antigen locus to autoimmune disease [7] and the association between variants at APOE and Alzheimers disease[8]. The success of these types of studies were limited and results were often not replicable, however, since they relied on candidate gene selection, they counted nominal significance ($p<0.05$) as definitive without controlling for multiple testing and they could not control for population sub-structure, which we now know to be a major confounder of association studies ("population stratification")[2].

The challenge to expanding beyond single genes was that it was not feasible to sequence every genome in a cohort of individuals to document all the variants therein. Fortunately, in the 1990s, a systemic genome-wide association approach to studying genetic variation associated to disease on a genome-wide scale without the need for sequencing was proposed that involved harnessing the LD previously described. A few key developments made such a proposal possible: (1) the discovery of widespread correlation between SNPs throughout the genome and the presence of a limited set of haplotypes at a given locus, which precluded the need to fully sequence someone's genome [9]; (2) the cataloging of common human genetic variation and the correlations therein in 2005 through the completion of The HapMap Project[10]; (3) the engineering of microarray technology – high-throughput arrays that facilitated cheaply testing hundreds of thousands of variants at a time[11]; and (4) the development of statistical methodology to analyze such large amounts of data and control for technical artifacts and biases therein as well as methodology for imputation of non-typed SNPs, which allowed for meta-analyses across many individual studies[12].

The proposal to study genetic variation on a genome-wide scale (GWAS: genome-wide association study) therefore involved focusing on a particular subset of variation: common SNPs. Common SNPs are defined as those whose minor allele frequency is > 1%. In the European population, there are 10 million such sites in the genome at which individuals' genotypes vary [13]. These sites comprise about 90% of an individual's heterozygous sites throughout their genome[2]. Practically speaking, this type of variation is extremely convenient because, as described, there is widespread correlation among common variants due to their being relatively old in evolutionary history and recombination happening mostly at hotspots. This means that only a subset of variants needs to be genotyped in a given study to serve as a proxy for the rest, and microarray technology can easily allow for cheap, direct genotyping of these hundreds of thousands (and now a million) SNPs.

Beyond practicality, however, there are theoretical arguments grounded in population genetics that predict the genetic architecture of common disease to be at least in part due to common variation (hence the so-called "common-variant common-disease" hypothesis). This argument includes the typical late onset of many common diseases that precludes causal alleles from strong natural selection, causal alleles being neutral in the past and only now having an effect due to recently introduced changes in living situations, recent population expansion allowing detrimental alleles to rise in frequency and phenomena such as heterozygote-advantage [2].

Perhaps more satisfying arguments for common variation playing a role in the etiology of common diseases are empirical ones. First, if the majority of causal variants for common disease were rare and highly penetrant, the linkage era would have seen

more success in mapping of common diseases (unless these diseases are typically caused by a single highly penetrant mutation in one of many possible genes, though so far there is no evidence for this model). Second, the genome-wide association era of the last decade has seen major breakthroughs in the identification of thousands of common variants associated to complex traits. Many dichotomous traits, such as Type 1 and 2 diabetes and inflammatory bowel disease, have seen a rapid expansion in the number of loci definitively associated to disease[14–21]. Likewise, huge success has been met by those studying quantitative traits in population cohorts, such as blood lipid levels, human height and body-mass index[22–26]. The National Human Genome Research Institute has cataloged over 5,000 common variants found to be associated to complex traits or diseases [27]. The loci discovered to date have identified common variants at previously known genes where variants of much stronger effect had been identified through linkage but also have identified numerous novel genes that implicate potential pathogenic mechanisms that would never have been suspected. In age-related macular degeneration, complement factor H was identified to bear common variation that predisposes to disease; in Crohn's disease, *ATG16L1* and *IRGM*, two genes known to play a role in cellular autophagy, are independently implicated; and regulatory variation near *SORT1* was found to predispose to higher blood LDL levels [16,28–30]. These are examples where GWAS has proven to be a powerful method to pinpoint genes and processes (and in the aforesaid cases specific variants) that are implicated in disease that would have otherwise been missed.

With the discovery of associated common SNPs, a few fundamental and for the most part universal observations have been made.

- **The effect sizes of risk variants are small.** While there are a cases where the effect size is observed to be ≥ 2 (such as *IL23R* in Crohn's[18], *SORT1* in blood-lipid levels[31], *APOE* in Alzheimer's disease[8] and *CFH* in age-related macular degeneration[32]), most lie between 1.1 and 1.5[2].
- **Complex traits are extremely polygenic.** A disease with substantial heritability and risk alleles individually of low effect is likely polygenic, i.e. phenotypes are contributed to by risk alleles of small effect at many genes throughout the genome. In Schizophrenia, for example, a significant inflation in association exists across thousands of common variant loci. Importantly, this inflation was replicated in family studies, precluding the possibility of it being driven by population stratification [33,34]. Subsequently, Yang et al. showed that much of the heritability for human height is explained by alleles of very weak effect that have yet to reach genome-wide significance[25]. Therefore, while the risk alleles discovered to date explain only a portion of the heritability, it may be the case for some complex traits that there are many more common variants of even lower effect that contribute to disease (though additional sources of variation not captured – rare variation, structural variation and epigenetic variation – are likely to contribute to the genetic architecture as well. Rare variation is discussed in section 1.2).
- **Most associated regions implicate many genes and have not yet been fine-mapped.** SNPs tag large regions of disequilibrium where continuous blocks of DNA that are between recombination hotspots are passed on together and rarely recombined. Thus, association at a particular SNP implicates not only the SNPs in

LD with the lead SNP but also any other type of variant nearby, such as another SNP or a CNV (as was the case with *IRGM* in Crohn's disease)[29]. These blocks can be 100s of kilobases in size, and contain up to 25 genes (median of 3 genes, distribution shown in Figure 1.1). In a very small number of cases, fine-mapping has led to the unambiguous identification of a causal variant; for the vast majority of associations discovered to date, however, the causal gene is not known (though in a few cases may be suspected). This limiting factor is what prevents the translation of these genetic discoveries to biological mechanism and ultimately to improving therapeutic design.

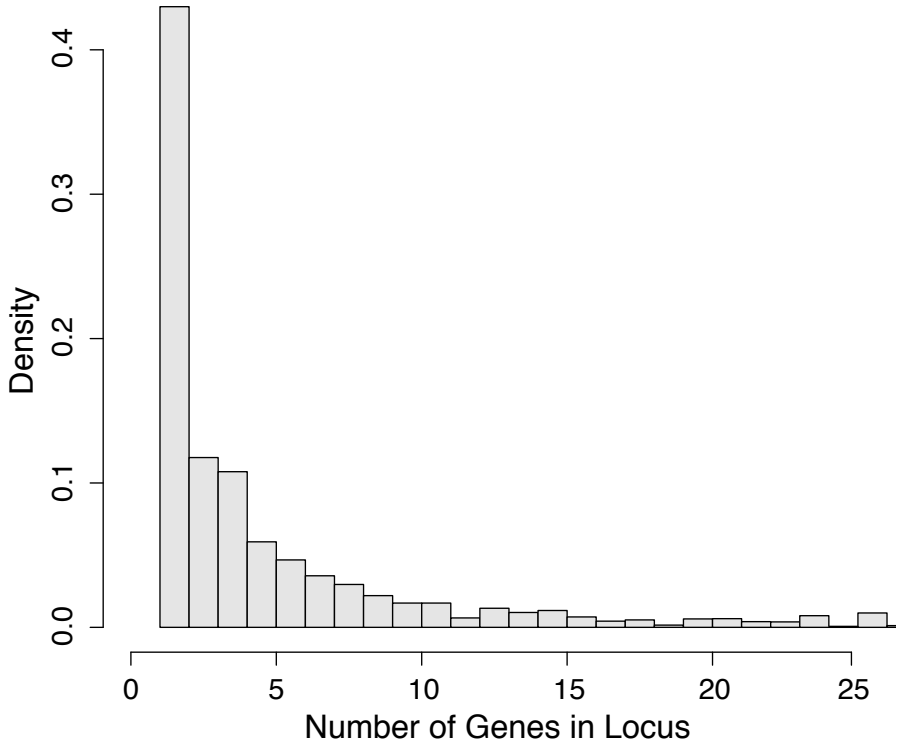


Figure 1.1 Distribution of number of genes per locus. 7,496 SNPs were downloaded from the NHGRI GWAS catalog, of which 5,924 were in LD with genes. The wingspan of a SNP was defined by the SNPs in $r^2 > .5$ of the lead SNP and further extended to the nearest recombination

Figure 1.1 (continued)

hotspots. Genes that overlap the defined wingspan were counted as part of the locus (see Chapter 2).

- **The same locus can be implicated by rare and common variants.** In many cases, common variants of modest effect were identified at the same locus that known highly penetrant mutations have been identified as causal of an extreme version of the common phenotype. For example, a SNP with minor allele frequency of 11% at *LDLR* has been identified in a study of blood-lipid levels, complementary to the more rare and deleterious mutation in the same gene known to cause familial hypercholesterolemia[31,35]; 4 independent common variants that have been found to be associated with variation in the cardiac QT-interval are near genes that are causal of the rare Mendelian long-QT syndromes [36,37]; and common variants near *FGFR4* were found to influence human height, where rare mutations in a similar growth-related gene (*FGFR3*) cause the Mendelian disease achondroplasia [38].
- **Most associations are not at coding variants but rather are in intergenic regions.** The majority of common variant associations that have been identified have no evidence of a nearby coding variant, suggesting that many genetic risk factors for common disease will interfere with gene expression rather than the protein itself[39]. This idea has been further supported by a number of expression quantitative-trait locus (eQTL) studies, where the allele of a SNP correlates with the expression of a gene in *cis* nearby, suggesting that it may interfere with the activity of transcription factors[40,41].

- **The genes implicated in associated regions fall into pathways.** Pathway analysis – the study of genes that co-occur in documented pathways or are connected in the biological literature – is an area of study that has been widely explored and has suggested that genes associated through GWAS tend to fall in cohesive pathways. Well-established methods include the GSEA, the Set-Screen test, MAGENTA, FORGE, ALIGATOR and INRICH[42–46]. Though distinct in their underlying methodology, each method tests a set of defined pathways or gene groups (using databases like KEGG, REACTOME, Gene Ontology, TargetScan, PANTHER and OMIM) for enrichment in association beyond random expectation, controlling for confounders like LD and gene size[6,47–49]. Though these approaches are inherently constrained by our limited knowledge of gene function and how genes should be grouped into pathways, they have nonetheless indentified potentially relevant biological pathways[50]. For example, cell adhesion molecules have been shown to be significantly enriched in associations to schizophrenia and bipolar disorder, pointing at synaptic formation as a potential biological lead underlying the etiology of schizophrenia[51].

There is a slowly growing set of instances where the preliminary GWAS scan has identified not only an associated locus but also a specific variant that has relevant biological findings and is likely driving the association signal. For example, the *ATG16L1* T300A variant associated to Crohn's disease was found to be the most associated SNP in the region and also suggested to be causal: *ATG16L1* is a gene involved in cellular autophagy (the removal of intracellular contents) and this particular

variant was shown to impair the process of selective autophagy associated with the removal of *Salmonella typhi* bacteria in human epithelial cells[30].

The vast majority of identified loci, however, do not point to specific causal variants or even causal genes. As shown in Figure 1.1, associated regions can contain multiple genes, making the problem even harder. With this in hand, it is clear that the next phase of GWAS should be focused on identifying important genes and pathways that are implicated by the genetic findings so far. However, while pathway analysis has promise, not all genes can be assigned to pathways and not all pathways are known. This thesis will discuss the use of protein-protein interaction data (which can be experimentally produced on a broad scale and without the requirement of pre-determined processes) to begin to elucidate some of the biological networks or complexes that underlie genetic associations to complex traits.

1.2 Exome sequencing and discovery of rare variants

Most GWAS arrays are not designed to capture rare alleles. As the identification of new common-variant loci with substantial effect size has tapered off for many traits, investigators have begun to wonder about rarer variation and are therefore turning to sequencing in order to capture this class of variant. A rare allele is defined as a minor allele frequency of <1% [52]. Most alleles in the population are rare, while common variants (though comprising most of the heterozygous sites of an individual) are relatively few. The etiology of this distribution is that variants enter the population spontaneously fairly rapidly (~40 *de novo* SNPs per individual[53]), but are quickly lost via genetic drift or are even more rapidly lost if they are deleterious. Some – both benign

and deleterious – will be passed on to progeny and persist at low levels before disappearing and thus populate the rare alleles present in humans, though very few will rise to a frequency considered common. The size of the existing pool of rare variants in the population has been further enhanced by the recent population expansion, which will naturally expand with it the number of rare alleles[52].

The reasons for testing rare variation for association to disease are many and have elicited lively debates among investigators. It is worth discussing three of the motivations behind this work.

- **Fine-mapping the causal gene in a common variant associated locus**. As discussed, common allele associations are hampered by long ranges of linkage disequilibrium, which leads to 10s-100s of SNPs being implicated under one signal and multiple genes to consider (Figure 1.1). It is possible to fine-map causal common alleles through conditional analysis (i.e., hold allele A fixed and test allele B to see if the signal from B is independent from A), but in regions where r^2 is close to 1 between associated SNPs, it can be impossible. Furthermore, common SNP associations do not necessarily implicate that a SNP nearby is the causal variant; SNPs can tag other types of genetic variation that may not have been captured such as CNVs or indels. On the other hand, since rare variation usually represents alleles that arose on a haplotype relatively recently, they are typically not in appreciable linkage disequilibrium with many surrounding SNPs. Therefore, if one can detect association to a rare allele (which is admittedly more difficult to do than for common alleles, see power discussion below), it is likely (though not always) that the rare allele itself – and not surrounding alleles in LD –

is the causal variant. If rare alleles affect the same gene in the region (i.e., the gene influences the phenotype through both common and rare variation in the population), this comes in handy: detection of rare associated alleles could help pinpoint the causal gene in a region of common-variant association [54].

- **Rare alleles are more likely to be deleterious**. As discussed, one of the reasons that an allele may be rare in the population is if it confers reduced fitness and therefore is not able to rise in frequency. One would then predict that these types of alleles – since their age precludes having been weeded out due to natural selection – may be enriched for deleterious events, and in fact it has been observed that rare CNVs are more likely to affect whole genes than common ones and rare SNPs are more likely to be non-synonymous than their common counterparts[52].
- **Heritability**. Though thousands of common variant loci have been discovered to be significantly associated across hundreds of traits, most of the results do not explain the majority of estimated heritability within each phenotype. Undoubtedly, alleles across the entire frequency spectrum will contribute to disease. The relative proportion of heritability explained by common versus rare alleles remains an open question and probably depends on the particular trait being studied. Nonetheless, it is possible that rare variation might fill some of the gap in heritability that many diseases in medical genetics currently face.

For the reasons discussed in section 1.1, the field has up until now focused on identifying association to SNPs whose allele frequency is > 1%. With the recent advent of better and more affordable sequencing technology, it is now possible to test for

variants below the 1% threshold. Many groups have focused initially on coding SNPs, since the functional consequence of these variants is easier to interpret and the alleles themselves likely carry a higher effect (which decreases the number of samples needed to assay). However, as the cost of sequencing decreases even further, groups will likely move to whole-genome sequencing, though the interpretation of rare, non-coding variants poses a new challenge.

Thus, investigators have embarked on targeted-, whole-exome- and sometimes whole-genome-sequencing studies. Targeted re-sequencing, wherein regions associated to disease through common variants are deeply sequenced, has the promise of fine-mapping the causal common variant and identifying low-frequency (and potentially greater effect) variants that may help fine-map causal genes. Whole-exome or whole-genome studies have the added benefit of identifying associated variants that point to novel regions/genes, especially for those traits for which fewer common variant associations have been documented (such as autism[55]).

Still, it is important to keep in mind the challenges that we face when looking for rare variants of low-effect. A common misconception is that through exome-sequencing, one will identify functional variants (i.e., missense, non-sense or splice-site) that will be easily interpretable as to their relevance to disease. On the contrary, in a sample of individuals, many neutral or disease-irrelevant rare coding variants will vastly out-number the alleles that might be relevant to disease. One promising avenue forward is to analyze these rare events as groups in the context of genes or groups of genes to improve our ability to identify causal variants. Chapter 4 will briefly review the existing rare-variant analysis strategies and will introduce the idea of integrating sequencing data with

protein-protein interaction data in order to find networks of proteins affected by rare variation that would have otherwise been missed by studying each variant separately.

1.3 Protein-protein interaction data

1.3.1 From association to disease mechanism

Whether studying common or rare variants, the true promise of human genetics lies in its ability to reveal novel disease mechanisms to be targeted by therapeutics. As discussed in sections 1.1 and 1.2, it is only for a subset of discovered GWAS loci so far that the causal gene is known, and an even smaller set where the implicated mechanism is known. It has previously been observed that genes causal of Mendelian diseases code for proteins that are part of the sample protein complex[56–59]. The hypothesis being tested in this thesis is that genetic variants predisposing to complex disease affect a common and limited set of cellular processes. Studying genes in the contexts of the biological networks through which they exert their effect is thus of utmost importance. Here, we rigorously test whether probing physical interactions among the products of associated genes offers a direct route to unraveling disease etiology.

1.3.2 Protein-protein interactions

Protein-protein interactions (PPI) represent the main functional units in the majority of cellular processes. They are the driving events in signal transduction, formation of macromolecular machinery, transportation of one protein by another to a different compartment of the cell, modification of proteins by phosphorylation, acetylation or ubiquitination, scaffolding of structures, and so on. The protein-protein

interaction is the ultimate work unit of almost all cellular processes. The scientific community has therefore built a field around discovering and documenting these interactions on a proteome-wide scale in a number of different model organisms and has been very successful in doing so, particularly in yeast but also in humans[47,60–66]. The hope is that these connectivity networks will provide a snapshot of the various processes going on in the cell.

The two main types of interactions are pair-wise interactions and macromolecular protein complexes. Each individual pairwise interaction can be furthermore categorized as stable or transient. Stable interactions are ones that persist over a period of time, such as a membrane scaffolding protein that organizes and clusters membrane ion channels, while transient interactions are ephemeral, such as a kinase phosphorylating a substrate. Typically, it is easier to capture stable interactions experimentally; therefore, most PPI experiments will be biased toward stable interactions. Macromolecular complexes are multiple proteins that bind to create a larger functional unit; RNA polymerase, comprised of multiple proteins that bind together to carry out transcription, is a classic example.

1.3.3 Protein-protein interaction experimental strategies

Different experimental approaches have been developed to capture PPIs. These can largely be broken down into genetic and biochemical approaches[67]. Genetic approaches typically refer to transcriptional complementation assays, such as yeast two-hybrid (Y2H), while biochemical approaches refer to affinity technologies, such as affinity purification followed by mass-spectrometry (APMS). As discussed in section 1.3.5 and shown in Figure 1.3, nearly 80% of the evidence for interactions in large PPI databases are from Y2H or affinity technology.

The Y2H method was original developed in 1989 by Fields and Song and was based on the observation that eukaryotic transcription factors have a modular structure, consisting of a DNA-binding domain and an activation domain responsible for recruiting transcriptional proteins such as RNA polymerase[68]. While different systems have been developed since its original inception, the yeast protein Gal4 construct is far-and-above the most common transcriptional complementation system used to date. Y2H involves transfecting yeast with plasmids that encode the two proteins of interest, each fused to one of to the two domains of Gal4. If the two proteins stably bind in yeast, the now-functional Gal4 recognizes a specific DNA sequence and activates transcription through recruiting RNA polymerase II [69]. The gene being transcribed can be used as a read-out, such as a His reporter gene.

The benefit of the Y2H system is that it is easily scaled up and may recapitulate the cellular milieu, especially when studying yeast proteins. However, when studying human proteins heterologously expressed in yeast, sources of false positives or false negatives include post-translational modifications that are different and may interfere with the interaction, additional proteins present in yeast that positively or negatively affect the interaction, inability of bait and prey to localize to the nucleus, poor folding of the bait and prey or needed proteins/molecules are not present[70]. The accuracy of the Y2H system has therefore been debated. An early study that compared a handful of interaction datasets to a gold standard of protein complexes discovered fairly low quality of the Y2H data[71]. In contrast, more recently, Y2H-reported interactions were deemed to be of high quality when compared to gold-standard sets of binary interactions, claiming that the previously reported comparison erroneously included protein complex

data [72]. Ultimately, Y2H should be interpreted as one source of interaction information that together with other approaches may contribute to a high-confidence set of interactions.

More recently, mammalian-2-hybrid systems have been developed to better approximate an in vivo setting for the tested interaction. These systems have extended the Y2H concept to two parts of a protein that together lead to a quantifiable signal – such as a split beta-glycosidase enzyme complementation system that leads to conversion of a substrate into a quantifiable product (such as X-gal, which turns blue after cleavage) [69]. The mammalian system relieves some of the problems in yeast, such as creating a more relevant cellular milieu.

The other main category of PPI experimental strategies is biochemical. *In vitro* biochemical experiments include phage display (display of a bait protein on a bacteriophage that then adheres to a prey protein immobilized on a surface), co-immunoprecipitation of recombinant proteins, or protein chips (immobilized recombinant proteins on a surface)[67]. Much like Y2H, these approaches usually assay binary interactions. *In vivo* (or less *in vitro*) biochemical approaches refer to the purification of endogenous protein complexes in cells. These typically use primary antibodies directed at the bait, ligands specific to the bait or affinity tags. Affinity purification (AP) refers to the process of purifying an endogenous protein of interest along with its interaction partners. Though there are many different flavors of AP, most are based on the purification of a protein from cell extracts (usually tissue culture but whole organs can also be used) along with its interaction partners[70]. Sources of variability in the different methods include the affinity protein – such as an antibody specific to the bait protein versus an antibody

specific to a peptide-tagged recombinant protein – and cell lysis conditions as well as whether the process is one-step or tandem[73]. The tandem method (Tandem Affinity Purification) employs a dual purification system whereby the protein is dual-tagged such that it can be purified twice in order to reduce the number of proteins purified non-specifically; one example is using a tag consisting of a protein A-calmodulin fused bait protein that will bind to IgG first followed by calmodulin beads and is very popular since all yeast open reading frames are available TAP-tagged[74]. In higher eukaryotes where these libraries are not available, HA-, FLAG- and Myc- tandem affinity purification is used though it requires over-expression of the fusion protein[75]. Since tandem purification is biased toward stable interactions, one-step purification techniques are ideally preferred because they will include weaker and more transient interactions. In addition, methods that do not require over-expression are preferred since over-expression can lead to non-physiologic interactions.

In AP-MS, quantitative mass spectrometry is then used after the affinity purification to indentify interaction partners. Mass spectrometry is traditionally used to identify the elemental composition of a sample, such as a purified protein and its binding partners. Quantitative mass spectrometry (q-MS) is a recent and critical advance in the field of proteomics: rather than simply knowing the identity of a molecule, q-MS uses stable isotope labeling of peptides or proteins in order to quantify their abundance. This technological advance has been revolutionary to the field of protein-protein interactions: it allows for the quantitative comparison between a specific bait and a control (such as a non-specific antibody or unrelated protein), which can then be used to rule out non-specific binding partners[76]. This precludes the need for extensive purification, which

had been previously required and resulted in the inability to perform high-throughput experiments. Additionally, with q-MS, one-step purification methods can now be employed that may identify weaker, more transient interactions than two-step approaches (such as TAP-tagging).

To ensure the capture of weaker, more transient interactions that would be missed in the washing steps of one- or two-step APMS, chemical cross-linking can be employed. Through the formation of covalent bonds, such as through the use of a His-Bio tag, weak interactions are stabilized and the protein of interest, along with its binding partners, can be purified[70].

Other methods include fluorescence-based approaches (such as Flourescence Resonance Energy Transfer), 2D gene-electrophoresis, synthetic lethality, and x-ray chrystallography, among others [77,78]. Ultimately, meta-databases (such as STRING and InWeb[62,79]) that collect multiple different sources of evidence to create a confidence score for each interaction are likely to be the most useful.

Table 1.1 Categories of protein-protein interaction experiments. Controlled vocabulary was downloaded from the IntAct database website[63] and the top two levels of the ontology are shown. Column 1 lists the broadest PPI categories and column 2 lists specific examples of each category.

Category	Examples
Biophysical	Neutron diffraction
	Light scattering
	Circular dichroism
	Intermolecular force
	Scintillation proximity assay
	Molecular sieving
	Neutron fiber diffraction
	isothermal titration calorimetry
	surface plasmon resonance
	Nuclear magnetic resonance
	Filter trap assay
	Mass spectrometry studies of complexes
	Small angle neutron scattering
	Electron resonance
	Amplified luminescent proximity homogeneous assay
	Electron diffraction
	X-ray crystallography
	Fluorescence technology
Genetic Inference	Random spore analysis
	Synthetic genetic analysis
Protein Complementation Assay	Transcriptional complementation assay
	3 hybrid method
	Cytoplasmic complementation assay
	Bimolecular fluorescence complementation
	Membrane bound complementation assay
Biochemical	Cross-linking study
	Comigration in gel electrophoresis
	Cosedimentation
	Footprinting
	Affinity technology
	Chromatography technology
	Enzymatic study
Imaging Techniques	Electron microscopy
	X-ray tomography
	Light microscopy
	Fluorescence microscopy
	Confocal microscopy
	Atomic force microscopy

1.3.4 Protein-protein interaction databases

There are over 25 publically available databases that collect published protein-protein interactions. Examples of such databases are YPD (the first to use this approach), MINT, BIND, HPRD, DIP, MIPS, IntAct, BioGrid, among others[60,61,63–66,80]. Almost all databases are built through literature curation, whereby expert curators sift through the scientific literature and report interactions and the controlled vocabulary that represents their experiment type (see Table 1.1 for an example of vocabulary). IntAct, one of the PPI databases, has articulated a very detailed ontology that generally represents the full spectrum of codes used by other databases[63]. There are ~50 different categories of experiments to detect PPIs, and many more sub-categories. Overwhelmingly, transcriptional complementation assays and affinity technologies comprise the majority of documented interactions. For example, in MINT, 86% of interactions are categorized as either "two hybrid pooling approach," "tandem affinity purification," "two hybrid," "anti-tag coimmunoprecipitation," "pull-down," or "anti-bait coimmunoprecipitation"[61].

These databases are incredibly powerful because they extract knowledge from the full body of published literature. Rather than having to invest time and money in building interaction networks *de novo* for each protein of interest, investigators can begin their search by studying what has already been done.

The drawback to these databases is that there is no weighting scheme for the reliability of interactions. As discussed in section 1.3.3, false positives can arise in Y2H experiments due to failure of yeast to provide the appropriate cellular milieu for eukaryotic proteins. Likewise, affinity technologies can result in false positives due to

non-specific binding of the affinity antibody to non-bait proteins. Furthermore, a recent study that compared these databases identified surprisingly little overlap among them[81]. As a result, groups have begun to address this issue by generating probabilistic meta-databases that pool PPIs and assign a confidence score to reflect the presumed reliability of the interaction.

1.3.5 InWeb Database

The meta-database that is used throughout this thesis is the InWeb database developed by Lage et al. in 2007. InWeb is a probabilistic database of reported protein-protein interactions [24,31] that combines interactions from MINT, BIND, IntAct, KEGG annotated protein-protein interactions (PPrel), KEGG Enzymes involved in neighboring steps (ECrel), Reactome, GRID, DIP, MPact, DOMINO, and HPRD [53-61]. All human interaction data were pooled, and to increase the coverage of interactions, interolog data (the transfer of protein interactions between orthologous protein pairs in different organisms) from the Inparanoid database were included by transferring from 17 eukaryotic organisms, similar to Lehner and Fraser[82,83]. The distribution of interactions categorized by species is show in Figure 1.2.

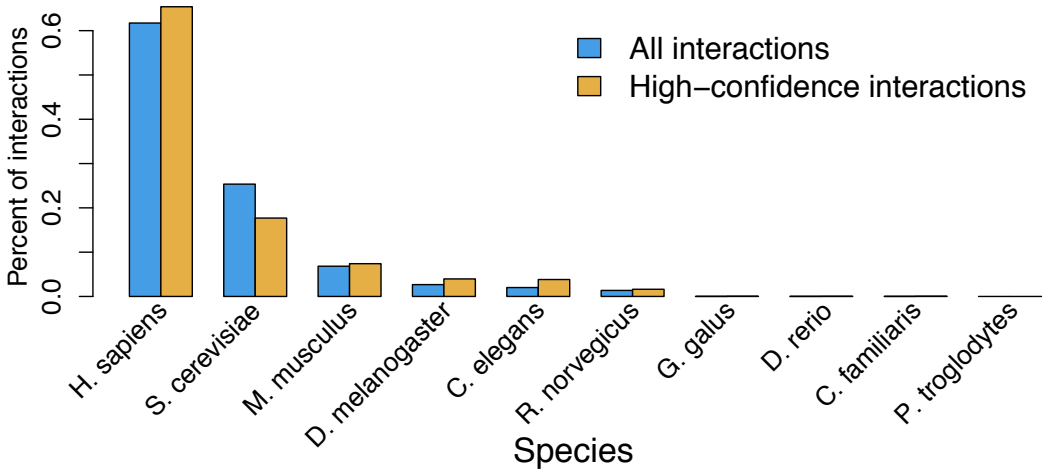


Figure 1.2 Distribution of interactions in InWeb categorized by species.

As mentioned, InWeb uses a probabilistic weighting scheme to assign each interaction a confidence score that combines three measurements: topology, scale and reproducibility. *Topology* refers to the sub-network surrounding the interaction of interest and was implemented using a topological scoring method described in de Lichtenberg et al. that assigns a raw score (RS) from zero to -∞ according to:

$$RS = -\log((NS_1 + 1) \cdot (NS_2 + 1))$$

where NS_1 and NS_2 are the number of non-shared interaction partners of proteins 1 and 2[84]. This score will be closer to zero as the number of non-shared partners decreases. *Scale* refers to the total number of interactions reported in the publication reporting the particular interaction of interest. Small-scale experiments were up-weighted because they typically provide multiple sources of evidence for an interaction, unlike large-scale screens [85]. Finally, *reproducibility* refers to the number of independent publications that reported the interaction, with the assumption that more independent reports of an interaction indicates better reliability. Topology, scale and reproducibility are combined using the following score:

$$Score = \frac{RS}{\sum_{i=1}^{N} 1/\log(i_{\text{int}})}$$

where i_{int} is the number of interactions reported in the ith publication for an interaction of interest. Confidence therefore increases as this score increases (approaches zero). This score was calibrated against 35,000 high-confidence human interactions and shown to reliably call these gold-standard interactions as high-confidence[62]. Version 3.0 of this database contains 428,430 interactions, 169,810 of which are deemed high-confidence, non-self interactions across 12,793 proteins. High-confidence is defined by a signal to noise threshold as determined by the calibration step.

The distribution of experiments that contribute to InWeb is shown in Figure 1.3. As expected from the majority of experiments that go into the smaller databases, affinity technology and protein complementation assays are the most abundant experimental evidence source in the database. The main difference between the entire set (blue bars) and the high-confidence set (gold bars) are that the high-confidence set has a higher protein-complementation to affinity ratio. Manual inspection of a subset of these publications reveals that often while the interaction is documented as Y2H, for example, it is typically a smaller publication with multiple lines of evidence for the interaction[62].

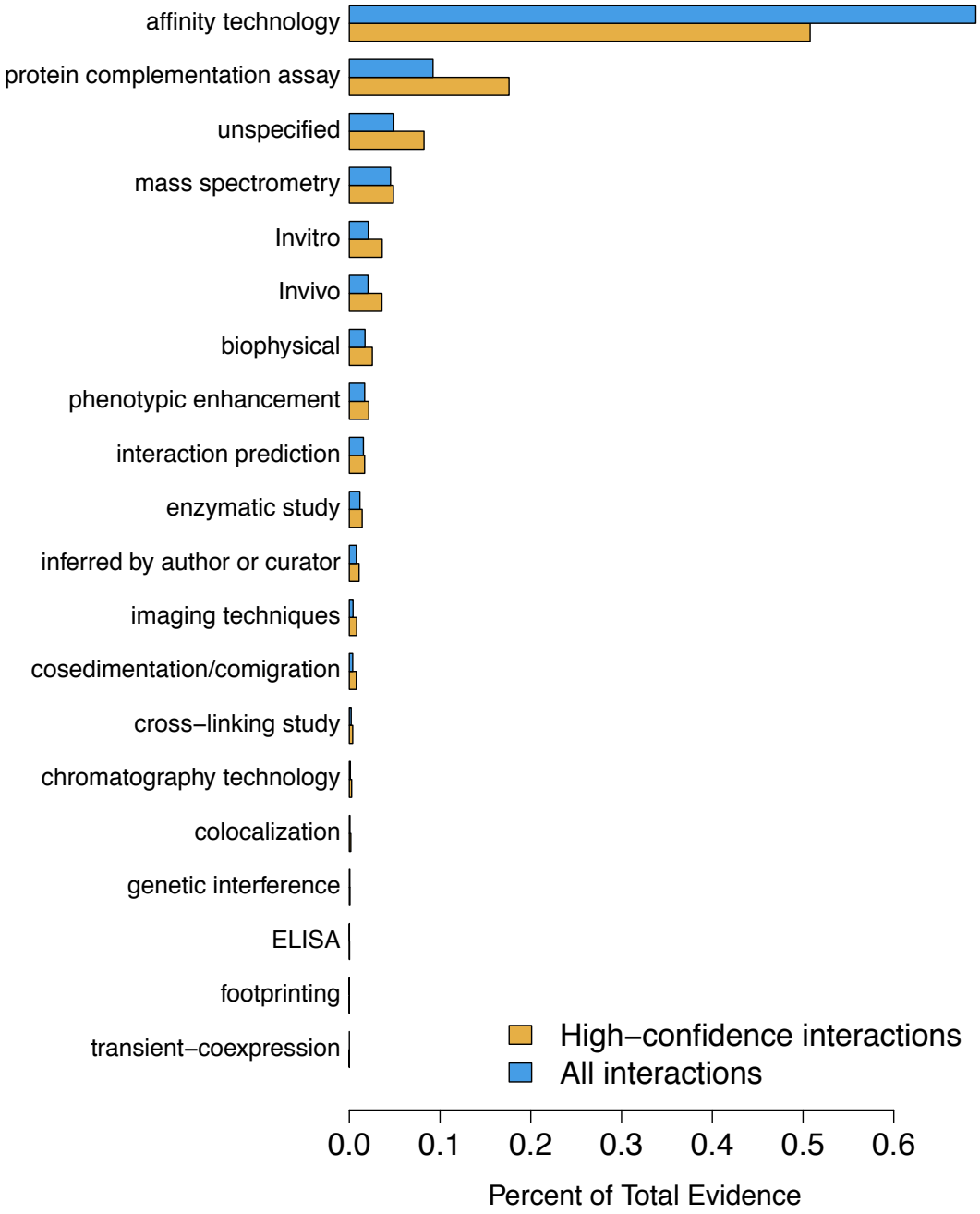


Figure 1.3 Distribution of experiment types in the InWeb database. Experimental evidence codes were gathered for all interactions in InWeb and grouped according to a broader category, if possible (typically evidence codes were similar to those listed in column 2 of Table 1.1).

Finally, the distribution of the number of independent publications per interaction for the InWeb database (blue bars) and the same data for the high-confidence set (gold bars) is shown in Figure 1.4. As expected, the high-confidence set has a higher mean number of publications per interaction.

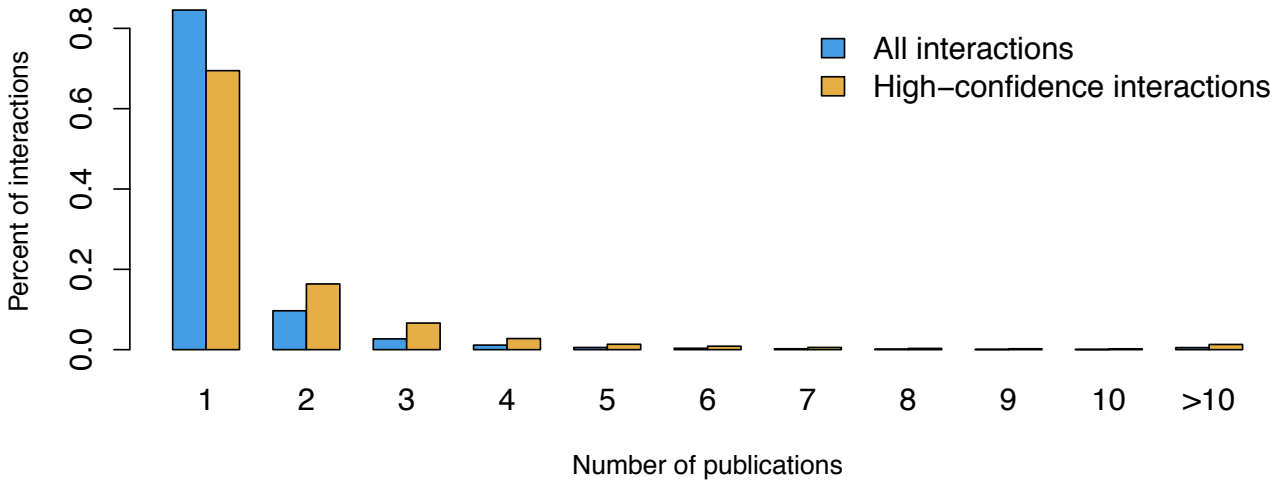


Figure 1.4 Number of publications per interaction in the InWeb database.

The binding degree distribution (see section 1.4) of the InWeb database is characteristic of biological networks in that there are many proteins with one or two interactions and few proteins with many interactions[86]. The distribution of binding degrees for proteins in the high-confidence set (which is the set used in future discussions) is shown in Figure 1.5. The database is very interconnected, as evidenced by an average clustering coefficient (C) of 0.261 (see section 1.4 for definition C)[87].

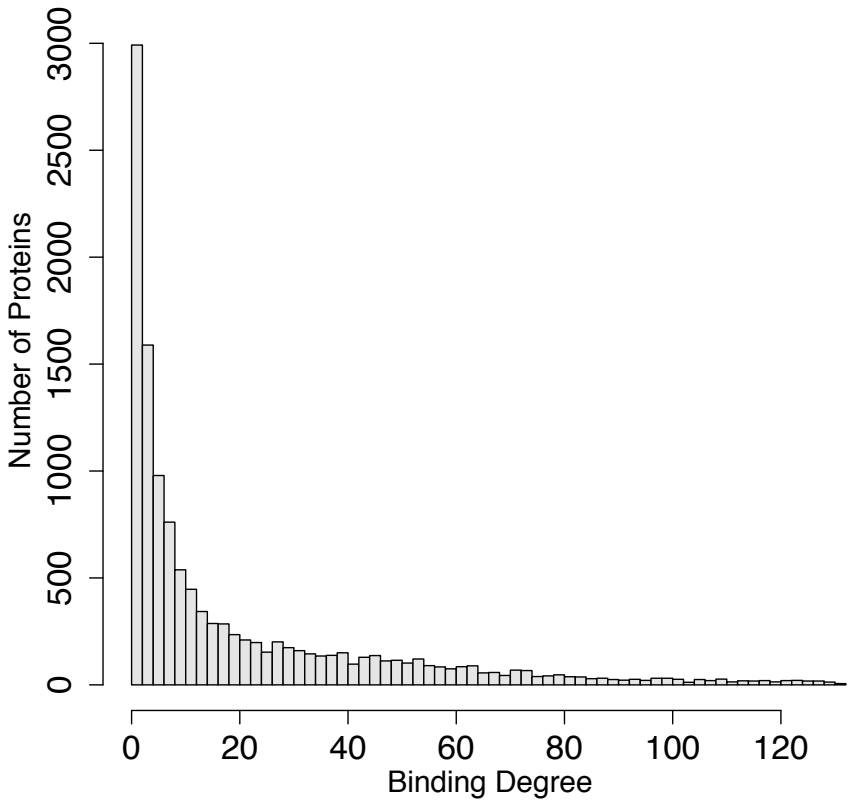


Figure 1.5 Binding degree distribution of InWeb high-confidence interactions. The number of high-confidence interaction partners was summed for each protein and plotted; outliers (>3 standard deviations away) were removed for plotting purposes.

1.4 Network analysis strategies: describing and evaluating networks

In this thesis, we combine lists of binary PPIs into networks, and it is therefore important to define the descriptive terms used to describe them. Note that the descriptions provided here are specific to this work but can take on more general meanings in graph theory.

The term "global network" refers to the entire PPI database, such as the InWeb database, which can also be represented as a large network. A sub-network refers to a subset of the global network. Sub-networks are usually what one observes – for example,

10 connected proteins of interest associated to autoimmunity represent a sub-network within the global network InWeb. The following terms are used throughout this thesis or in the field of network analysis to describe networks:

- **Network, graph, database:** These words are used interchangeably, though "database" is specifically a global term.
- **Node**: A protein in the network
- **Edge**: Evidence of two proteins interacting. Networks can be represented with weighted or binarized edges, depending on whether the strength of interaction is relevant. Binarizing networks eases the calculation and interpretation of many network statistics, and therefore this thesis will use binary edges. Additionally, edges can be unidirectional (as in an enzyme acting on a substrate) or bidirectional (a ligand binding its receptor). In this thesis, we will consider all interactions to be bidirectional.
- **Binding degree**: The number of edges incident at a protein in a network.
- **Average clustering coefficient**: The cliquishness of a network, i.e. the average probability that a node's binding partners bind each other. In a protein-protein interaction database, this is typically high.
- **Diameter**: The farthest distance between nodes in the network.
- **Shortest-path**: For two nodes of interest, the fewest number of edges between them.
- **Power-law**: The term used to describe a degree distribution that follows $P(k) = k^{-\gamma}$ where $\gamma>0$. Typically, biological networks follow the power-law (as opposed to a poisson degree distribution, which a classic random network will follow).

- **Sub-graph, sub-network**: A subset of the global database.
- **Size**: The number of nodes in a network.
- **Connectivity**: A broad term used to describe the connectedness of networks. Various statistical measurements on networks can be referred to broadly as connectivity.

Typically, these are the terms used to describe both global and sub-networks, although many of them (average clustering coefficient, diameter, degree distribution) are used more often to describe global networks. It is important to note that it is not readily clear which measurements are best to use, or that a sufficient set of terms exist to describe a network[88]. The relevance of the terms listed here are that they comprise important characteristics to match if generating randomized networks for comparison. In Chapter 2, we will define additional measurements specific to sub-networks.

If we can describe a network with the aforesaid terms, why do we need to analyze it? Network analysis is a broad term that refers not only to determining metrics used to describe networks but also to methods to determine the significance that should be applied to local areas of connectivity. In the context of this thesis, the reason to evaluate networks (rather than simply describe) is to determine which ones we should care about. For example, for the sub-network built from 10 autoimmune proteins, what values of connectivity (such as the number of connections) are significantly above chance expectation? Network connectivity will behave like any other statistic: there exists a null distribution (albeit one that is not straightforward to estimate) that is expected simply from random networks, and network analysis can be used to estimate this null distribution and assign significance to sub-networks.

When evaluating whether a sub-network is significantly interconnected, the goal is to compare it to randomized, matched networks. The earliest randomized graph model is the Erdös-Rényi random graph, which is simply the set of edges between randomly drawn pairs of nodes[89]. In 1959, Gilbert then proposed the model E=pN(N-1)/2, where E is the number of edges, p is the probability of two nodes connecting, and N is the number of nodes. Another model, the "geometric graph," is built by distributing points randomly in space and connecting any two points that are below a set distance ε apart[89]. These three random graphs allow for analytic calculations of expected connectivity. Permutation approaches, on the other hand, build random graphs by permuting the existing database. *Edge shuffling* refers to random edge permutation with degree preservation (i.e., randomly permute edges but keep each protein's binding degree the same)[86]. *Node randomization* refers to randomly drawing nodes from the larger database to create a random graph. Variants on node randomization preserve degree by sampling nodes of the same degree distribution as the ones in the observed sub-network.

The described approaches to random graph generation each fail to recapitulate many of the global properties of real protein-protein interaction networks. Both the ER and geometric graph will have a different global degree distribution and average clustering coefficient, and the Gilbert approach also has a different degree distribution. Global edge-shuffling will indeed match the degree distribution but will have a much lower global average clustering coefficient; random node selection will fail to match the properties of the observed network, since random nodes are not likely to share the same degree distribution. Node randomization that specifically selects nodes matched on observed degree distribution is theoretically ideal but practically constrained, since there

is usually a limited number of possible permutations that will match the observed degree distribution. Chapter 2 introduces a novel permutation scheme to randomize networks matched on degree distribution and average clustering coefficient that is not limited by permutation possibilities.

1.5 Summary

The hypothesis being tested here is that both common and rare genetic variation associated to disease affect a common and limited set of cellular processes, and that introducing PPI data to genetic association results will help identify networks underlying risk variation. So far, we have broadly introduced efforts to identify common variants associated to disease through GWAS, recent sequencing efforts to search for rare variation, protein-protein interactions and the experiments used to generate them and finally how PPI databases are built and analyzed. Chapters 2-4 of this thesis will delve more deeply into historical methods to integrate genetic and PPI data as well as introduce novel statistical approaches to integrating PPI data from both public databases as well as direct experiments with genetic data from GWAS and sequencing studies.

In chapter 2, we discuss the design, testing and implementation of a novel *in silico* approach ("DAPPLE"[87]) to rigorously ask whether loci associated to complex traits code for proteins that form significantly connected networks. In chapter 3, we study protein complexes associated to variation in the electrocardiographic QT-interval, a heritable phenotype that when prolonged is a risk factor for cardiac arrhythmia and sudden cardiac death. In chapter 4, we consider whether PPIs can be used to interpret rare and *de novo* variation discovered through recent technological advances in exome-

sequencing. Ultimately, we hope that this thesis convinces its readers that considering genetic variants in the context of the biological networks through which they exert their effect is a promising step forward in complex trait genetics.

2 Disease Association Protein-Protein Link Evaluator

The contents of this chapter also appear in *PLoS Genetics* as:

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7(1): e1001273. doi:10.1371/journal.pgen.1001273

Individual contributions are listed at the end of the chapter.

2.1 Abstract

Genome-wide association studies (GWAS) have defined over 150 genomic regions unequivocally containing variation predisposing to immune-mediated disease. Inferring disease biology from these observations, however, hinges on our ability to discover the molecular processes being perturbed by these risk variants. It has previously been observed that different genes harboring causal mutations for the same Mendelian disease often physically interact. We sought to evaluate the degree to which this is true of genes within strongly associated loci in complex disease. Using sets of loci defined in rheumatoid arthritis (RA) and Crohn's disease (CD) GWAS, we build protein-protein interaction (PPI) networks for genes within associated loci and find abundant physical interactions between protein products of associated genes. We apply multiple permutation approaches to show that these networks are more densely connected than chance expectation. To confirm biological relevance, we show that the components of the networks tend to be expressed in similar tissues relevant to the phenotypes in question, suggesting the network indicates common underlying processes perturbed by risk loci. Furthermore, we show that the RA and CD networks have predictive power by demonstrating that proteins in these networks, not encoded in the confirmed list of disease associated loci, are significantly enriched for association to the phenotypes in question in extended GWAS analysis. Finally, we test our method in 3 non-immune traits to assess its applicability to complex traits in general. We find that genes in loci associated to height and lipid levels assemble into significantly connected networks but did not detect excess connectivity among Type 2 Diabetes (T2D) loci beyond chance.

Taken together, our results constitute evidence that for many of the complex diseases studied here, common genetic associations implicate regions encoding proteins that physically interact in a preferential manner, in line with observations in Mendelian disease.

2.2 Introduction

Common genetic variants in over 200 genomic loci have now been unequivocally associated to immune-mediated diseases by genome-wide association studies (GWAS)[15–17,19,20,23,90–97]. It is presumed that these associations represent perturbations to a common but limited set of underlying molecular processes that modulate risk to disease. The next challenge – and the great promise of human genetics – is the identification of these disease-causing pathways so they may be targeted for diagnostics and therapeutic intervention.

In identifying such processes, there are difficulties in both (i) identifying the specific genes at (and how they are molecularly impacted by) each association and (ii) inferring disease-causing mechanisms from the set of identified genes. Linkage disequilibrium blocks containing disease-associated SNPs can be hundreds of kilobases in size, and some contain tens of genes to consider. Genes are often informally implicated in pathogenesis by their proximity to the most associated marker, their biological plausibility, or simply their being the only protein-coding gene in the region. In reality, however, it is only a very small subset of confirmed GWAS associations for which specific functional variants have been proven experimentally.

More systematic approaches have been applied to connect genes to a common process with the use of independent data, such as Gene Set Enrichment Analysis (GSEA) and Gene Relationships Across Implicated Loci (GRAIL) [45,90,98,99]. Both approaches identify connections between genes based on descriptive categories that outline the theorized underlying pathogenesis. However, these concepts are often general, so that specific hypotheses and molecular pathways can be difficult to define and are somewhat limited to established knowledge bases.

Observations of interactions between the products of protein-coding genes offer the most direct route to identifying pathogenic processes. It has been shown in a number of Mendelian diseases that genes causal of a particular phenotype tend to physically interact [62,100–102]. This has been confirmed in the model organism *C. elegans*, where RNAi knock-down of germline genes correlated highly with their products interacting in yeast-two hybrid experiments[102]. A classic example of a human Mendelian disease that recapitulates this model is Fanconi Anemia (FA), an autosomal recessive disorder linked to at least 13 loci, at least 8 of which function together in a DNA repair complex [100]. Protein-protein interaction (PPI) data has also been used to formulate hypotheses about co-expressed genes as well as cancer genes [103,104]. We note that previous attempts to use PPI data to prioritize candidate genes in Mendelian disorders have been successful as was the case with the published tool *Prioritizer* [59]. We therefore set out to test such an approach in complex disease.

Investigators have rapidly populated databases of such protein-protein interactions over the past decade. The reported interactions in PPI databases stem from both small, directed investigations and high-throughput experiments, primarily yeast two-hybrid

screens and affinity purification followed by mass spectrometry [105]. These data are inherently noisy: beyond technical false positives and negatives, experiments *in vitro* may report interactions that do not occur *in vivo* simply because the proteins involved never overlap spatially or temporally. To mitigate the noisiness of PPI databases, we extract networks from "InWeb", assembled in 2007 by Lage et al [62]. InWeb is a database of 169,810 high-confidence pair-wise interactions involving 12,793 proteins (human proteins and their orthologs). Lage et al. define high-confidence interactions as those seen in multiple independent experiments and reported more often in lower-throughput experiments [62]. To further restrict the data to biologically plausible interactions, we overlay mRNA expression information to confirm co-expression of binding partners; this correlates with co-localization, similar phenotype and participation in a protein complex [106,107].

Assessing the significance of networks built from PPI data is challenging for two reasons: first, overall connectivity is a function of the binding degree (number of connections in the database for a given protein) of proteins within the network. Thus, the apparent density of a network could simply be due to the lack of specificity with which its constituents bind *in vitro*. Second, certain processes are more extensively studied, so more connections between proteins involved in them may be reported (see Figure 2.1; immune proteins are reported in more publications and have a higher mean binding degree). This confounds our effort to assess connectivity of associated loci if there is a bias in the data. From a genetic standpoint, a common randomization method would involve sampling SNPs from the genome matched for the appropriate parameters (such as

gene density and protein binding degree). This method becomes highly limited if the disease loci contain genes that are better studied than the randomly sampled SNPs.

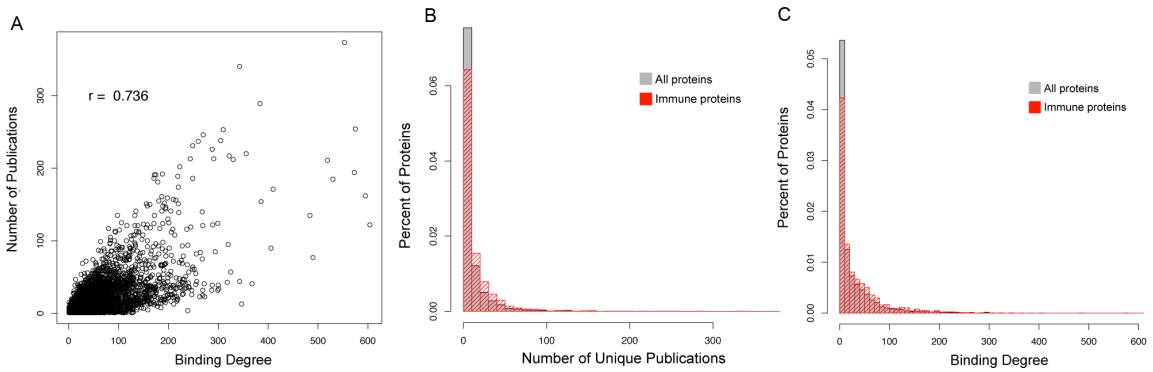


Figure 2.1 Immune proteins comprise a disproportionate portion of the PPI literature. For each protein, we enumerated the number of unique publications in which it is reported. Using a publically available expression dataset of 55 immune tissues[108], we then categorized proteins as "immune" or "non-immune". Genes in the top 2% of expression for at least 2 of the 55 immune tissues were deemed immune genes; their proteins were then deemed immune-proteins. The distribution of publications for immune proteins is significantly more than the background distribution (Kolmogorov-Smirnov p-value <2E-16).

Therefore, we apply a permutation method that is robust to non-specific binding and differences in publication density. We perform a within-degree node-label permutation that is carried out as follows: a random network is built that has nearly the exact same structure as the original InWeb network, only the node labels (i.e. the protein names) are randomly re-assigned to nodes of equal binding degree; this method assumes a null distribution of connectivity that is entirely a function of the binding degree of individual proteins. Random networks will have the same size, number of edges and per-protein binding degree as InWeb; we build 50,000 different random networks. With this

method, we are able to test the non-randomness of an observed sub-network *conditional* on the exact binding degree distribution of the observed disease proteins.

Others have used PPI data in complex disease to understand epistatic loci or to build a network of interacting proteins from associated loci [109–111]. The novelty of our method lies not in the idea that PPI data can be used to help understand genetic loci associated to disease, but rather in that we have developed a broadly-applicable method to statistically evaluate the degree to which non-random PPI networks emerge from loci associated to complex disease and to leverage from this insight about causal proteins in large loci [109,110]. We show this to be the case in a number of diseases.

Here, we use this methodology to evaluate whether genes in loci associated to five complex traits are significantly connected via protein-protein interactions. We report an algorithm to build and assess PPI networks using the InWeb database and find robust, statistically significant networks underlying associations to RA, CD, height and lipid levels, which we suggest as representative of the underlying pathogenic molecular processes. We then perform several detailed analyses on the RA and CD networks to confirm that they contain true biological insight into disease. We use independent mRNA expression data to show that the prioritized associated proteins we propose as interacting are co-expressed in relevant immune tissues, supporting a plausible biological setting for our findings as well as the validity of the reported protein-protein interactions. Lastly, by analyzing more recent GWAS meta-analysis results, we show that these networks contain components that show significant evidence of further genetic associations: proteins interacting with multiple associated network members and encoded elsewhere in the genome themselves carry an excess of association to disease in the latest

meta-analyses of each of these diseases. Our method, available for download, generates an experimentally tractable hypothesis of the molecular underpinnings of pathogenesis.

2.3 Methods

2.3.1 Network construction and evaluation pipeline

We construct and evaluate networks of disease loci as outlined in Figure 2.2. We first define *associated proteins* as gene products encoded in genomic loci harboring variants associated to disease (Figure 2.2A-B; see Materials and Methods for locus definition). We construct networks of protein-protein interactions representing proteins as nodes connected by an edge if there is *in vitro* evidence of interaction (InWeb high-confidence interaction set). We build direct networks, in which any two associated proteins can be connected by exactly one edge, and indirect networks, where associated proteins can be connected via *common interactor* proteins (not encoded in associated loci) with which the associated proteins each share an edge. We restrict direct and indirect interactions to only those between proteins encoded in distinct associated loci.

We then calculate several metrics to evaluate network properties. These metrics can be divided into two categories: an edge metric and node metrics. The edge metric is the *direct network connectivity* parameter defined as the number of edges in the direct network. We interpret *direct network connectivity* as the frequency with which different loci harbor proteins that directly bind each other, regardless of how they assemble; *direct network connectivity* is therefore our most straightforward metric. Node metrics include the following: *associated protein direct connectivity* and *associated protein indirect connectivity* which refer to the number of distinct loci an associated protein can be

connected to directly and indirectly, respectively, and *common interactor connectivity* which refers to the average number of proteins in distinct loci bound by common interactors in indirect networks. We interpret all three node metrics as descriptive of the type of network that was constructed: a stream of connections (such as the network A-B-C-D-E) will likely have low and insignificant node metrics despite a significant edge metric, whereas a more tightly clustered network might be enriched for both edge and node metrics. We assess the statistical significance of the various connectivity parameters using a within-degree node-label permutation strategy that controls for variation in the degree to which certain proteins are studied or behave *in vitro* (Figure 2.2C). As we are interested in the processes underlying disease, we also define the gene encoding the top-scoring protein in each locus as most likely to be causal for association (Figure 2.2D). We then use tissue expression data to test whether our nominated candidate genes are enriched in the same tissue(s) and therefore participate in a network that is biologically feasible (Figure 2.2E). With this approach, we aim to construct plausible models of biological networks underlying pathogenesis.

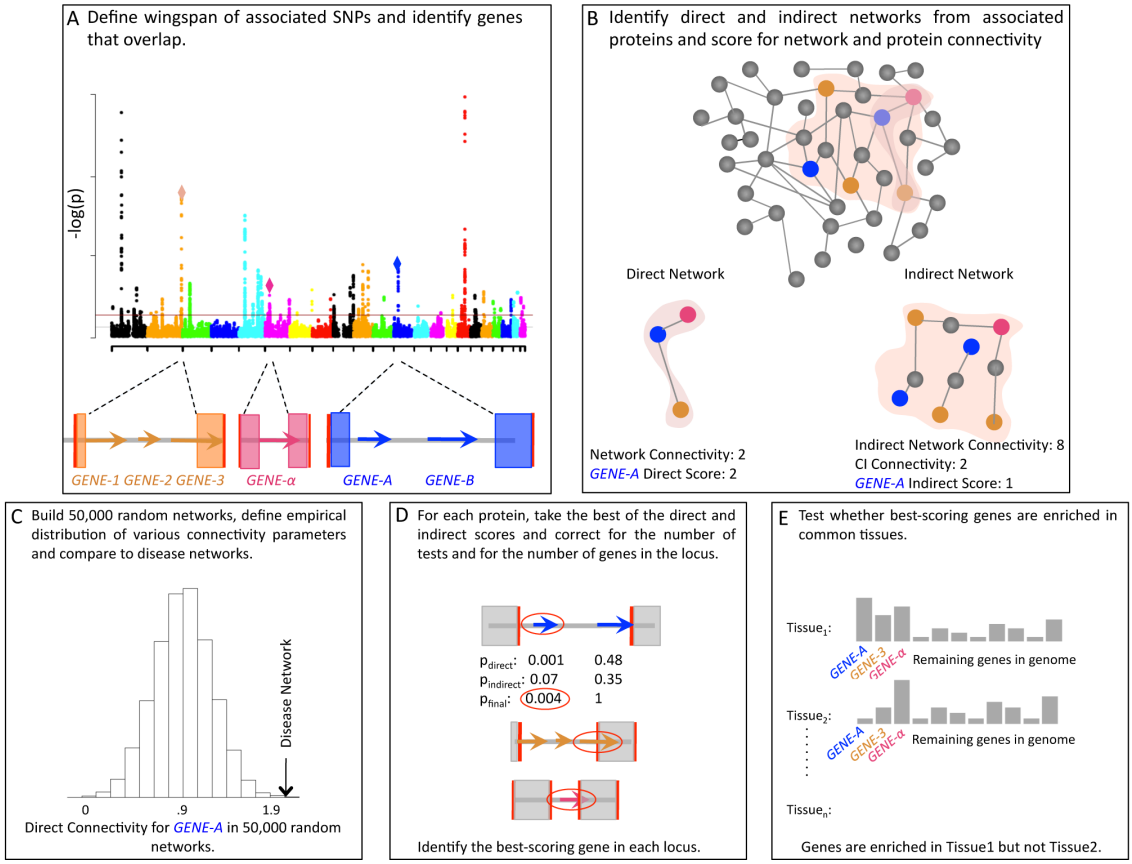


Figure 2.2 Pictorial outline of methodology. A. Genes overlapping the wingspan of associated SNPs are defined, and these genes code for *associated proteins*. **B.** Associated proteins are used to recover direct and indirect networks. Direct networks (left) are built from direct interactions between associated proteins according to the InWeb database (colored proteins). Connections between proteins within the same locus are not considered. Indirect networks (right) are built by allowing connections between associated proteins through a protein elsewhere in the genome (grey). Various network parameters to quantify connectivity, defined in the text, are assigned. **C.** Random networks are built from a within-degree node-label permutation method. An empirical distribution is constructed for each network parameter and used to evaluate the significance of networks. **D.** Using the same permutation method to score individual proteins, a subset of proteins per locus is nominated as candidates for harboring causal variants (red circles). Scores used to nominate candidates are Bonferroni corrected for the number of possible candidates

Figure 2.2 (continued)

within each locus. **E.** Candidate genes from D (nominal p-values used) are tested for co-expression.

2.3.2 Network analysis

The goal of building random networks is to compare disease networks to what is expected given the binding degrees of associated proteins. In order for the random networks to represent appropriate comparisons, they must mimic the original networks in their overall structure; furthermore, the binding degree of any one protein in a random network should be close to its value in the original network. As was discussed in section 1.4, many of the existing algorithms to generate random networks do not achieve this matching. As such, we used a within-degree node-label permutation method and built random networks whose network topology, as measured by clustering coefficient and conformity to the Power Law, is close to the original network and whose individual proteins have the same degree of binding as in the original network [86,112].

We define the original PPI network G (the InWeb network) to have a node corresponding to each protein known to participate in a protein-protein interaction, where an edge represents such interaction. Let G have n vertices and E edges. Let G_0 be a random graph with the same set of nodes as G and randomly-assigned edges. Moreover, for every node i_0 in G_0 $\deg(i_0) = \deg(i)$ where $\deg(i_0)$ is the binding degree of node i_0. The algorithm for generating a graph G_0 from a given graph G involves two steps, a permutation step and a switching step.

Permutation Step: Let $k_1, k_2, k_3, ..., k_m$ be the sequence of all possible node degrees in G and let K_i be the set of vertices that have degree k_i with $1 \leq i \leq m$. The following procedure permutes sets of same degree nodes 1,000 times:

1. Repeat the following 1,000 times:

2. For every set K_i, $1 \leq i \leq m$

    (a) Choose randomly two vertices a and b.

    (b) Swap its positions in set K_i.

The permutation step has a high impact in randomizing graph G but it has one limitation: the algorithm cannot perturb nodes that have unique degrees. The presence of high degree nodes in the network, often referred to as "hubs," that have unique degrees creates several situations where our method cannot permute. Let G_{unique} be the union of sub-networks consisting of high-degree nodes and their edges. In order to completely randomize graph G we apply an edge permutation algorithm for network G_{unique}.

Switching Step: The edge permutation algorithm starts from a given network and involves carrying out a series of switching steps whereby a pair of edges (A−B, C−D) is selected at random and the ends are exchanged to give (A−D , B−C) or (A−C , B−D). The exchange is only performed if it generates no multiple edges or self-edges. The entire process is repeated some number E times, where E is the number of edges in G_{unique}. The switching algorithm is used as a second step in our method, continuing the randomization process of graph G after the permutation step. It is applied only to G_{unique} graph defined by the set of nodes with unique degrees and their edges. The following procedure perturbs G_{unique}.

While there are nodes unvisited:

(a) Choose randomly edges A−B and C−D

(b) If A−D and B−C do not exist

i. Add edges A−D and B−C to G_0

ii. Remove edges A−B and C−D from G_0

The benefit of the permutation method used is that we repeat the entire process to generate 50,000 permuted networks matched for size, binding degree of proteins within it and overall network structure. As we show in the candidate gene section, this method importantly allows us to score individual proteins in the network, in addition to the network as a whole.

2.3.3 Evaluation of Permutation Method

To test whether this permutation method created random networks that matched the InWeb network in overall properties, we computed the binding degree distribution and the clustering coefficient for each random network. The binding degree distribution of the InWeb network, which is scale-free, follows the power law, and the random networks should too if they are structured like the InWeb network [62,86,106]. Because we permute within-degree, the distribution should be identical; indeed, we found that the random networks follow the power law and their distribution, which fits $k^{-1.7}$ (where k is equal to a given binding degree), are equivalent to the InWeb network. Secondly, the average clustering coefficient, C_{avg}, which is the probability that two binding partners of a vertex are connected, was computed for all nodes in the random networks and the InWeb

network. We found that the InWeb average clustering coeffecient was close to that of the random networks: C_{avg} for the InWeb network was 0.261 and the mean C_{avg} for the random networks was 0.197. The difference is due to the small amount of edge permutation that we perform on nodes with unique degrees.

Next, we tested the distribution of p-values that the permutation method reported for randomly selected groups of 30 SNPs (to roughly mimic RA and CD loci sets). If the permutation method correctly tested the null hypothesis, then the distribution of p-values for each metric should be uniform. For 50 such random groups of SNPs, we find via a χ^2 test for uniform distribution that the *network connectivity*, the *common interactor connectivity*, and the *associated protein indirect connectivity* fit what is expected under a uniform distribution ($p = 0.923$, $p = 0.787$ and $p = 0.896$). The *associated protein direct connectivity* is skewed towards non-significant p-values ($p = 1$), and thus the p-value distribution for this metric is less uniform. However, the skew towards p=1 for random SNPs indicates that for this metric, the permutation method would be more likely to call a false negative rather than a false positive.

2.3.4 Nomination of candidate genes within loci.

We applied an iterative scoring method to nominate candidate genes in multigenic loci. The goal of the approach is to identify a subset of candidate genes per locus (preferably one candidate gene although risk variants, if regulatory, could feasibly affect multiple genes) that are more highly connected to disease loci than by chance via permutation, or that score the highest compared to other proteins in the locus.

For a given gene in a multigenic locus, we identify whether it participates in the direct network only, the indirect network only, or both. If it participates in the direct network, we enumerate the number of distinct loci it connects to, D, and compare this number to the values obtained in permuted networks D_i for the i^{th} permutation. The number of successes, S, is enumerated, where S=1 if D_i is greater than or equal to D, otherwise S equals 0. After 50,000 permutations, the direct score for that protein is therefore:

$$\frac{\sum_{i=1}^{50000} S\Big|_{0:D_i<D}^{1:D_i\geq D}}{50000}$$

Equation 2.1

If the protein participates in the indirect network, we perform a similar enumeration. A caveat to the indirect connections is that unlike direct connections, a protein can indirectly connect to another protein in multiple ways and to a locus in even more ways. Based on the biological assumption that more indirect connections suggests more relatedness (functionally speaking), we would like to up-weight additional connections; as such, the indirect binding score I between a protein and another locus is the maximum number of indirect connections to a protein in that locus over all proteins in that locus. We compare I to the values I_i obtained in permuted networks for the i^{th} permutation. The number of successes, S, is enumerated, where S=1 if I_i is greater than or equal to I, otherwise S equals 0. After 50,000 permutations, the indirect score for that protein is therefore:

$$\frac{\sum_{i=1}^{50000} S\Big|_{0:I_i<I}^{1:I_i\geq I}}{50000}$$

Equation 2.2

If a protein participates in both networks, the direct and indirect scores are Bonferroni corrected for two tests and the best score is assigned.

Thus, each protein emerges with a final score that is used to nominate candidate genes within a locus; this score is further Bonferroni corrected for the number of possible candidates in that locus. In the text we distinguish between genes that achieve a corrected $p < 0.05$ from those that only achieve nominal significance. Nominal significance refers to significance after correcting for 2 tests (where applicable) but without correction for the number of genes in a locus; however, when building final networks (Figure 2.8) we use nominally significant genes. Genes in single-gene loci are automatically nominated if they participate in either network, but are only included as "candidates" if they achieved $p < 0.05$. This process is iterated until convergence where upon genes scoring $p < 0.05$ are nominated as the definitive causal gene, and all other genes in that locus are removed for the next iteration.

To validate that p-values are comparable across proteins, we tested for a correlation between prioritization p-value and binding degree. Such a correlation would indicate that the network significance assigned to individual proteins is mainly a function of representation in InWeb rather than relevance to disease processes. First, we selected 25 sets of 30 random gene-containing SNP wingspans, built random networks and ran

them through the pipeline using 5,000 permutations. We then evaluated the p-values assigned to individual proteins (1046 proteins in total across all random networks). The R^2 between binding degree and –log(p) was 0.000094 (p=0.757). We then collected the scores for proteins from all 5 complex traits discussed here (408 proteins in total) and R^2 was 0.0065 (p=0.104). These results are shown in Figure 2.3. We therefore conclude that p-values assigned to proteins are not confounded by the degree to which they are represented in the database.

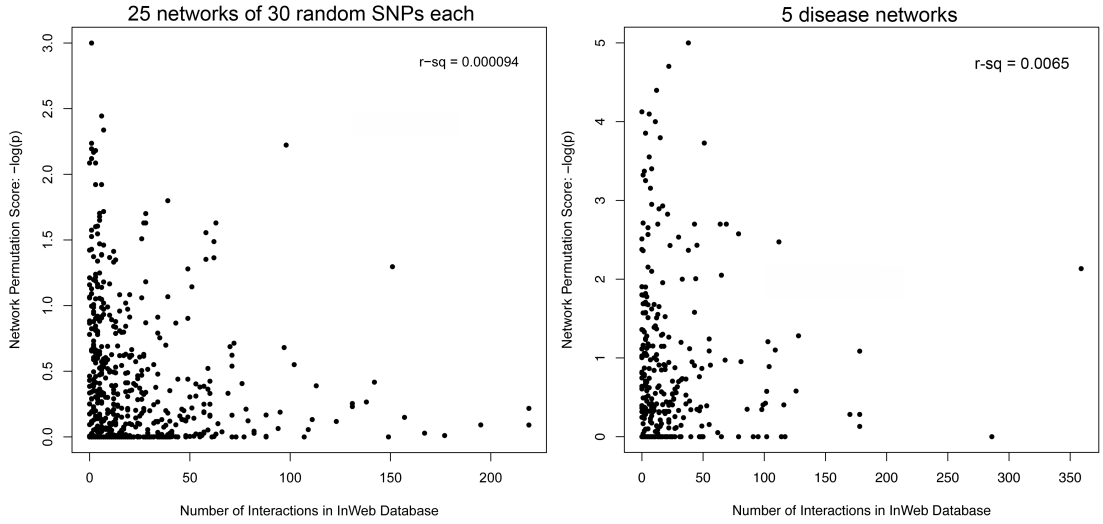


Figure 2.3 Correlation between prioritization p-value and binding degree. We show that there is no correlation between prioritization p-values given to genes and their representation in the database. We show this in randomized networks (**A**) and in the disease networks (**B**).

This analysis pipeline, which we call Disease Association Protein-Protein Link Evaluator (DAPPLE), is available for download at http://www.broadinstitute.org/mpg/dapple.

2.4 Results

2.4.1 Gene products encoded in associated loci interact

We first tested DAPPLE on the Mendelian disease Fanconi Anemia (FA) as a proof of principle. We input 9 of the FA genes and found 23 connections among them; compared to 50,000 random networks, the FA network is enriched for connectivity (*direct network connectivity* $p << 2\times10^{-5}$, Figure 2.4). The *associated protein direct connectivity, associated protein indirect connectivity* and *common interactor connectivity* were all significantly enriched ($p < 1 \times 10^{-5}$, $p = 0.00150$, $p = 0.00373$, respectively). These results agree with the current understanding of FA pathogenesis[56]. The network is shown in Figure 2.4.

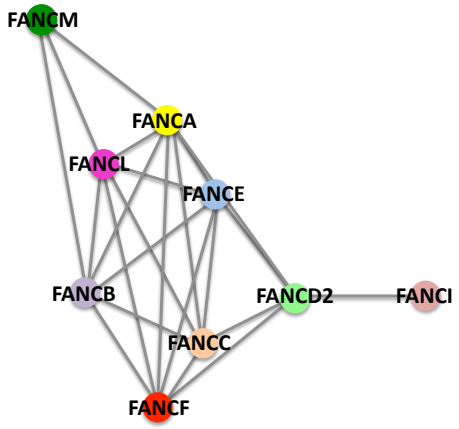


Figure 2.4 Fanconia Anemia network. As a benchmark analysis, we tested the method on Fanconia Anemia, a Mendelian disorder caused by genes that code for interacting proteins. 9 of the 13 FA genes were in the InWeb database. We found that the direct network connectivity was 23, which is many more than expected by chance (p<<2E-5). The *associated protein direct*

Figure 2.4 (continued)

connectivity, *associated protein indirect connectivity* and *common interactor connectivity* were all significantly enriched ($p<2E-5$, $p = 0.004$, $p = 0.009$, respectively). These results agree with the current understanding of FA pathogenesis. FA Network is shown.

We then set out to test our method on two autoimmune diseases that are both complex traits. Recent GWA studies in autoimmune and inflammatory diseases have been particularly successful at determining loci encoding risk to disease, with over 100 loci described to date [15–17,90,92,93]. We investigated rheumatoid arthritis (RA) and Crohn's disease (CD) and built networks from proteins encoded in 25 and 27 gene-containing associated loci, respectively. As described above, we built direct and indirect networks for each set of loci, evaluated the significance of the 4 network metrics to assess the probability that such networks could arise by chance, and we nominated candidate genes by assessing network participation. We followed up our results by assessing tissue co-expression as a test for the biological feasibility.

We were able to connect 20/27 loci for RA and 12/25 loci for CD in direct networks, strongly suggesting functional connections between proteins encoded in the associated regions. When compared to 50,000 random networks, we found that the *direct network connectivity* (the number of direct network edges) was statistically significant (27 for each disease; $P_{RA} = 3\times10^{-4}$, $P_{CD} = 1.11\times10^{-3}$; Figure 2.5) as was the *associated protein direct connectivity* (Figure 2.5, $P_{RA} = 0.02$, $P_{CD} = 0.00305$). Thus disease-associated loci encode directly interacting proteins beyond chance expectation, suggesting that risk variants may act on suites of proteins involved in the same process.

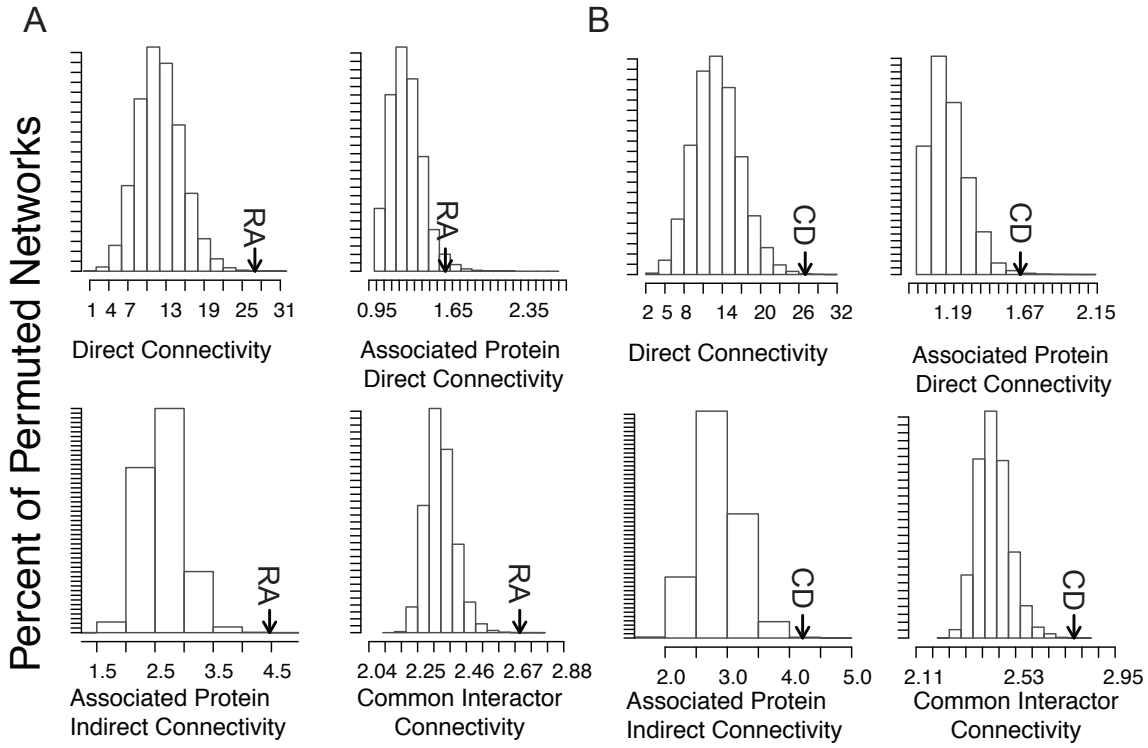


Figure 2.5 RA and CD direct networks are significantly interconnected. The *direct network connectivity*, *associated protein direct connectivity, associated protein indirect connectivity* and *common interactor connectivity* were enumerated for the disease networks and 50,000 random networks. A histogram is plotted to represent random expectation, and the disease network is shown by an arrow for **A.** RA and **B.** CD. From left to right and top to bottom, the connectivity p-values are: RA - 0.00031, 0.02000, 2.9734E-5, 6.9380E-5; CD - 0.00111, 0.00336, 0.00023, 0.00014.

We were then able to connect all but one gene-containing associated loci in each disease by expanding our networks to include common interactors (26/27 in RA; 24/25 in CD). The *associated protein indirect connectivity* was significantly enriched in both

diseases (Figure 2.5, $p = 3.0\times10^{-5}$ in RA, $p = 2.3\times10^{-4}$ for CD), as was the *common interactor connectivity* (Figure 2.5, $p = 7 \times 10^{-5}$ for RA and $p = 1.4\times10^{-4}$ for CD).

Our approach controls for biases in the data: using the high-confidence interactions from InWeb addresses laboratory artifacts, and node-label permutation accounts for ascertainment biases due to differing levels of knowledge on biological processes for those proteins present in InWeb (Figure 2.1). We show empirically that priority scores given to proteins have no correlation with the degree to which they are represented in the database (Figure 2.3). A fundamental limitation of any functional data is that genes for which data are missing will not be considered. This applies to similar methods, including expression data that can be limited to genes represented on specific arrays or ontology analyses that are restricted to well characterized genes. Here, proteins that are entirely absent from the filtered InWeb data are not considered in our analysis (see Discussion). It is important to note that these genes cannot be ruled out as potentially affected by causal variation since we have no power to make such a conclusion. We note, however, that the loci we have considered here (for the 5 complex traits) have the majority of their genes present in the high-confidence InWeb database (median inclusion of 81.5%).

2.4.2 Alternate network analysis

To ensure the robustness of our significance, we tested two alternative methods: we (1) built networks from randomly selected SNPs and (2) permuted all the edges, rather than only networks of nodes with unique degrees (see section 1.4 for discussion of network randomization methods). We carried this out in CD and RA. Permuting the

SNPs requires that the randomly chosen loci be matched for gene content as well as average binding degree of encoded proteins; method 1 is thus severely limited by the strict matching criteria, making this method unsuitable, and additionally, it does not easily allow for scoring of individual proteins. Thus, we permuted 1000 times and remove permutations for which the binding degree distribution of proteins in randomly selected loci was different (binding degree was greater or less than the mean of the disease proteins' binding degrees plus or minus 10, respectively, and the protein with the highest binding degree was more than the disease protein with the highest binding degree plus 20). Method 2 involves randomly shuffling the edges such that the number of edges per protein is preserved but the identity of binding partners is changed. Overall, the significance is replicated, though the SNP matching to a lesser extent: for RA, the *direct network* connectivity, *associated protein direct connectivity, associated protein indirect connectivity and the common interactor connectivity* p-values were p = 0.005, = 0.013, = 0.07, and = 0.06, respectively. For CD, the same p-values were p = <0.001, = 0.003, <0.001, and = 0.033. In the case of edges shuffling, the RA p-values were p = <0.001, = 0.013, <0.001, and <0.001 while the CD p-values were p = <0.001, <0.001, <0.001, and = 0.001. While we were encouraged to see persistent significance, SNP permutation may not be robust in the presence of extremes of gene density or protein binding degree at some loci, and edge permutation does not preserve the network structure of InWeb.

2.4.3 RA and CD networks identify new proteins enriched for association

In aggregate, these results suggest that the observations of connectivity in Mendelian diseases are recapitulated in both RA and CD and that common risk variants

predisposing to these diseases may impact sets of interacting proteins. Given the significant connectivity of common interactors in the indirect networks for RA and CD, we speculated that common interactors might themselves be affected by previously undescribed risk variation. To test this, we consulted association data for each disease in the available data from meta-analyses, which for RA was in a newly completed meta-analysis and for CD was the same study that yielded the 30 loci[16,113]. We assigned each recombination hotspot-bounded linkage-disequilibrium (LD) block in the genome an association score that represents the maximum score in that block corrected for the number of independent SNPs therein using logistic regression. Genes were assigned association scores based on the blocks they overlap; this score distribution can then be compared to the scores of all gene-containing blocks in the genome (for both diseases, we removed the MHC from this analysis due to LD properties). Using this method, we found that common interactors expressed in the same tissues as associated proteins in our networks (see below) were encoded in regions with evidence of association significantly in excess to what is expected in gene-containing regions. In RA, the distribution of common interactor scores was skewed toward higher association (one-tailed rank sum $p = 1.7\times10^{-5}$) and in CD, we saw similar enrichment ($p = 6.5\times10^{-4}$). This observed skew suggests that the common interactors themselves may harbor risk variants; we therefore considered the regions they overlap as candidates for replication (see section 2.4.6).

2.4.4 Extending Analysis to Height, Lipids and Type 2 Diabetes

To test whether the observed significant connectivity seen in RA and CD was present in non-immune complex traits, we tested our method on three traits: human

height, blood lipid concentration (both LDL and HDL) and Type 2 Diabetes (T2D). We used 37 replicated gene-containing loci associated with human height, 18 with blood lipid levels and 36 with T2D [19–24,95–97]. The loci associated to height and lipids each contain proteins that assemble into significantly connected direct networks (Figure 2.6, *direct network connectivity* $p = 1\times10^{-4}$ and $p = 1.9\times10^{-4}$ for each disease, respectively; see Figure 2.6 for significance of other 3 parameters). In the height network, 19/42 loci participated in the direct network and 34/42 participate in the indirect networks, but only the *direct network connectivity* and the *common interactor connectivity* were significantly greater than chance. In the lipids network, 11/19 participated in the direct network and 16/19 in the indirect; all node metrics except the *common interactor connectivity* were significantly enriched. 9/37 T2D loci participated in the direct network and 34/37 in the indirect network; however, 3/4 metrics were not greater than chance expectation and only one was slightly enriched (Figure 2.6, *direct network connectivity* $p = 0.44960$; see Figure 2.6 for significance of other 3 parameters).

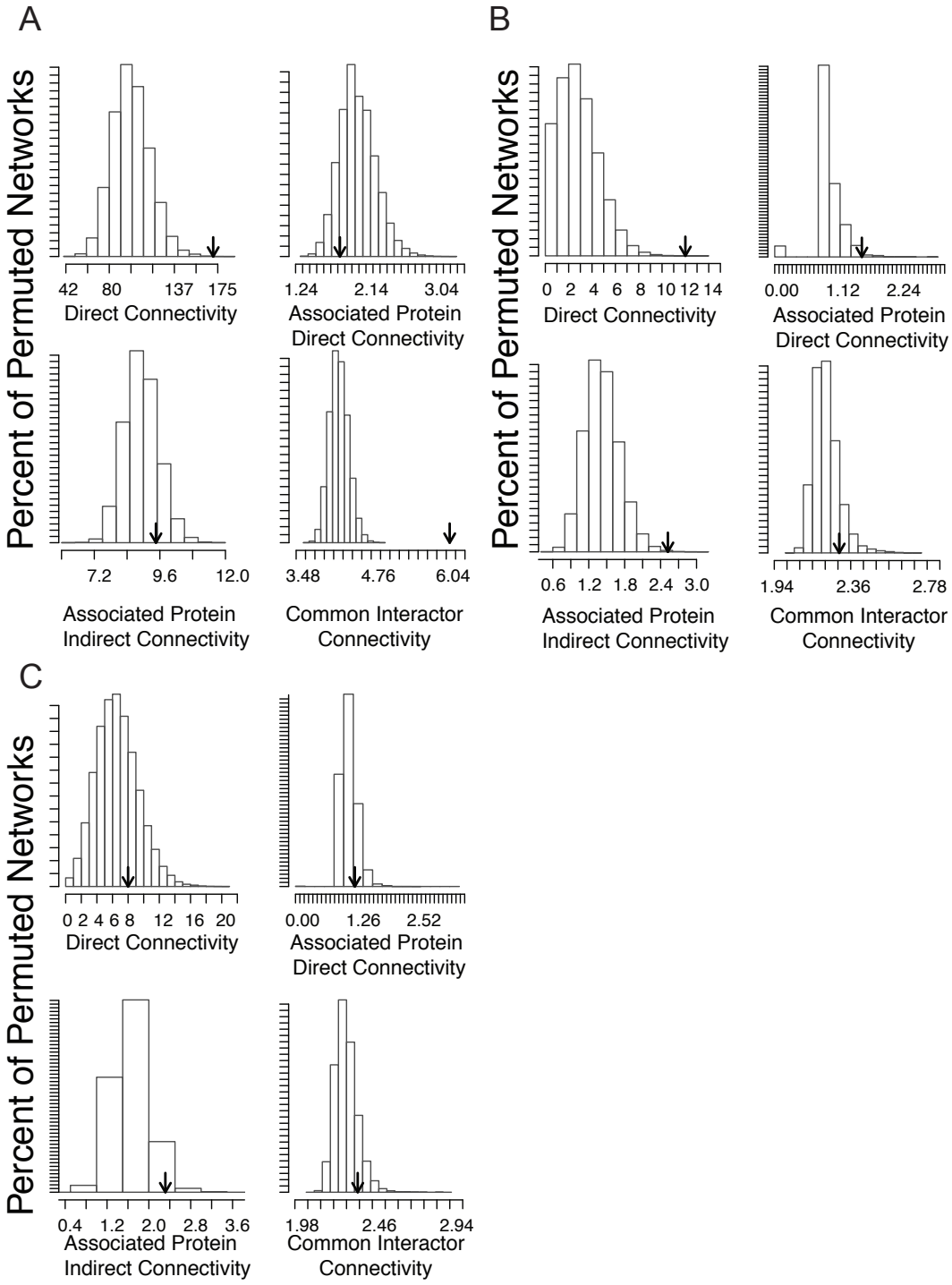


Figure 2.6 Network parameters for Height, Lipids, and T2D. We show the expected random distribution in relation to the disease network (arrow) for each of the 4 parameters (*direct network*

Figure 2.6 (continued)

connectivity, associated protein direct connectivity, associated protein indirect connectivity,common interactor connectivity) for height (**A**), lipids (**B**), and T2D (**C**). Permuted networks were generated use the within-degree node-label permutation method. Connectivity parameter scores are as follows, from left to right and top to bottom. Height: 1e-04, 0.8446, 0.192, <2E-5. Lipids: 0.00018, 0.01810, 0.00092, 0.13537. T2D: 0.41698, 0.23713, 0.03202, 0.2371.

We therefore conclude that the PPI connectivity seen in two autoimmune diseases can be generalized to other complex trait loci (height- and lipid-associated regions), though we could not confirm the significance of the T2D network.

Our results suggest that functionally connected proteins reside in regions of the genome associated to disease risk. Permutation analysis revealed that these connections are in excess compared to what is expected given the binding profiles of associated proteins. For RA and CD, other proteins interacting with the associated proteins also show evidence of association beyond chance expectation. Cumulatively, these findings suggest that risk to the complex disease/traits studied here is spread over functional groups of proteins, directly analogous to observations in Mendelian traits.

2.4.5 Prioritizing proteins in associated loci reveals likely pathogenic tissues

An obstacle to interpreting GWA results stems from the difficulty in identifying the genes within associated regions influenced by risk variants. Candidate genes are often selected based on proximity to most associated markers and miscellaneous forms of

previous knowledge. We therefore asked whether our observations could lead us to a principled, data-driven approach to selecting candidate genes by assessing their role in our networks. As shown Figure 2.2 and described in detail in section 2.3.2, we used an iterative optimization method to assign priority scores to associated genes based on the network participation of their encoded proteins. We nominate genes that achieve the best score within their locus as the candidates for influencing disease risk. We describe the results in detail here for RA and CD; see Table 2.1 and Table 2.2 for scores assigned to RA and CD.

We were able to nominate candidate genes in 12/21 RA loci encoding multiple genes (Table 2.1). Examples of candidate genes in RA were *IL2RA, CD40, CD28, PTPN22, CTLA4 and TRAF1*. We accomplished the same task in CD, nominating candidate genes in 10/18 multi-genic loci. Candidates included *JAK2, STAT3, IL23R/IL12RB2, PTPN2, MST1R* and *AIRE*. For both diseases, genes in single-gene loci are also scored, though they are automatically considered the candidate gene (but not necessarily part of the underlying mechanism). It is important to note that we do not expect high-scoring proteins in every locus; we only expect high scores for those proteins that may participate in the common process(es) detected via enrichment in connections. RA and CD, like most complex diseases, most likely have many underlying processes, not all of which are captured here.

Table 2.1 RA candidate genes proposed through permutation. Each protein that participated in the direct and/or indirect network was assigned a permutation p-value corresponding to the likelihood of seeing the degree of connectivity observed by chance. A protein's score was

Table 2.1 (continued)

Bonferroni corrected for the number of genes in its locus and for two tests if it participated in both the direct and indirect network ($P_{nominal}$ reflects the first correction). For tissue enrichment analysis and plotting of networks, we used the nominal p-value. Here, the permutation process is iterated twice, and the second iteration removes proteins in a locus that score $p > 0.05$ if any protein in that locus scored $p < 0.05$. Columns from left to right: SNP, number of genes in locus, gene, nominal p-value, corrected p-value, second iteration nominal p-value, second iteration corrected p-value. *NA indicates that the gene was not included in the second iteration, either because it was filtered after the first or because its participation depended on a protein that was filtered. Second iteration scores are only used to nominate candidate genes if no gene in the locus achieved $p<0.05$ the first time around.

Rheumatoid arthritis						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs3218253	2	IL2RB	1.6E-04	3.2E-04	1.0E-05	2.0E-05
rs2476601	6	PTPN22	1.4E-04	8.4E-04	1.7E-05	1.0E-04
rs10919563	1	PTPRC	4.0E-05	4.0E-05	2.0E-05	2.0E-05
rs3087243	1	CTLA4	8.0E-05	8.0E-05	1.0E-04	1.0E-04
rs4750316	1	PRKCQ	2.0E-06	2.0E-06	2.0E-04	2.0E-04
rs1980422	2	CD28	5.6E-04	1.1E-03	2.3E-04	4.7E-04
Rs540386	4	TRAF6	7.3E-03	2.9E-02	6.7E-04	2.7E-03
rs11586238	1	CD2	1.3E-03	1.3E-03	1.0E-03	1.0E-03
rs3761847	6	TRAF1	2.9E-03	1.7E-02	1.9E-03	1.1E-02
rs2812378	2	DCTN3	2.2E-03	4.4E-03	2.1E-03	4.1E-03
rs12746613	3	HSPA6	4.2E-03	1.3E-02	2.5E-03	7.4E-03
rs4810485	4	CD40	3.7E-03	1.5E-02	3.0E-03	1.2E-02
rs7574865	2	STAT1	8.9E-03	1.8E-02	4.3E-03	8.6E-03
rs12746613	3	FCGR2A	7.0E-03	2.1E-02	5.9E-03	1.7E-02
rs3766379	4	CD48	1.6E-02	6.2E-02	1.3E-02	5.2E-02
rs2395175	11	HLA-DRB5	2.1E-02	2.0E-01	1.5E-02	1.5E-01
rs2395175	11	HLA-DQA2	2.1E-02	2.1E-01	1.5E-02	1.6E-01
rs2395175	11	FKBPL	1.7E-02	1.7E-01	1.6E-02	1.6E-01
rs7574865	2	STAT4	2.8E-02	5.5E-02	2.0E-02	4.0E-02
rs5029937	1	TNFAIP3	2.2E-02	2.2E-02	2.1E-02	2.1E-02
rs3890745	4	HES5	2.7E-02	1.0E-01	2.3E-02	8.9E-02

Table 2.1 (continued)

Rheumatoid arthritis						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs2395175	11	HLA-DQA1	4.4E-02	3.9E-01	3.5E-02	3.2E-01
rs2395175	11	HLA-DQB1	5.0E-02	4.3E-01	4.1E-02	3.7E-01
rs2104286	4	IL15RA	5.7E-02	2.1E-01	5.1E-02	1.9E-01
rs1678542	20	MARS	9.2E-02	8.6E-01	5.8E-02	7.0E-01
rs6822844	3	IL21	6.3E-02	1.8E-01	6.0E-02	1.7E-01
rs2104286	4	IL2RA	7.6E-02	2.7E-01	6.3E-02	2.3E-01
rs2395175	11	PBX2	7.1E-02	5.6E-01	6.5E-02	5.3E-01
rs1678542	20	ARHGAP9	9.1E-02	8.5E-01	7.1E-02	7.7E-01
rs1678542	20	DCTN2	0.12	0.93	0.08	0.80
rs6822844	3	IL2	0.10	0.27	0.09	0.24
rs3890745	4	PEX10	0.09	0.32	0.09	0.31
rs13031237	2	REL	0.13	0.23	0.10	0.19
rs1678542	20	CDK4	0.12	0.93	0.12	0.92
rs1678542	20	CTDSP2	0.18	0.98	0.17	0.97
rs1678542	20	SLC26A10	0.23	0.99	0.21	0.99
rs1678542	20	AGAP2	0.23	0.99	0.22	0.99
rs10865035	1	AFF3	0.29	0.29	0.26	0.26
rs3761847	6	PSMD5	0.30	0.88	0.27	0.85
rs13031237	2	PEX13	0.30	0.51	0.28	0.49
rs3766379	4	CD244	0.32	0.79	0.31	0.77
rs2736340	2	BLK	0.35	0.58	0.33	0.55
rs12746613	3	FCGR3A	0.38	0.77	0.34	0.71
rs3761847	6	GSN	0.38	0.94	0.35	0.92
rs548234	2	ATG5	0.41	0.65	0.37	0.61
rs3890745	4	TNFRSF14	0.43	0.90	0.39	0.86
rs2395175	11	AGER	0.42	1.00	0.41	1.00
rs540386	4	RAG1	0.43	0.89	0.41	0.88
rs2736340	2	GATA4	0.73	0.93	0.71	0.92
rs4810485	4	NCOA5	0.86	1.00	0.84	1.00
rs1678542	20	PIP4K2C	0.14	0.95	1.00	1.00
rs1678542	20	DDIT3	1.00	1.00	1.00	1.00
rs1678542	20	INHBE	1.00	1.00	1.00	1.00
rs1678542	20	METTL1	1.00	1.00	1.00	1.00
rs1678542	20	GLI1	1.00	1.00	1.00	1.00
rs1678542	20	R3HDM2	1.00	1.00	1.00	1.00
rs1678542	20	XRCC6BP1	1.00	1.00	1.00	1.00
rs1678542	20	DTX3	1.00	1.00	1.00	1.00
rs1678542	20	INHBC	1.00	1.00	1.00	1.00
rs1678542	20	OS9	1.00	1.00	1.00	1.00

Table 2.1 (continued)

Rheumatoid arthritis						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs1678542	20	TSFM	1.00	1.00	1.00	1.00
rs1678542	20	AVIL	1.00	1.00	1.00	1.00
rs1678542	20	KIF5A	1.00	1.00	1.00	1.00
rs2104286	4	RBM17	1.00	1.00	1.00	1.00
rs2104286	4	PFKFB3	1.00	1.00	1.00	1.00
rs2395175	11	NOTCH4	1.00	1.00	1.00	1.00
rs2395175	11	AGPAT1	1.00	1.00	1.00	1.00
rs2395175	11	HLA-DRB1	1.00	1.00	1.00	1.00
rs2395175	11	GPSM3	1.00	1.00	1.00	1.00
rs3761847	6	RAB14	1.00	1.00	1.00	1.00
rs3761847	6	FBXW2	1.00	1.00	1.00	1.00
rs3761847	6	C5	1.00	1.00	1.00	1.00
rs3766379	4	LY9	1.00	1.00	1.00	1.00
rs3766379	4	ITLN1	1.00	1.00	1.00	1.00
rs3890745	4	MMEL1	1.00	1.00	1.00	1.00
rs4810485	4	MMP9	1.00	1.00	1.00	1.00
rs4810485	4	ZNF335	1.00	1.00	1.00	1.00
rs540386	4	RAG2	1.00	1.00	1.00	1.00
rs540386	4	FLJ14213	1.00	1.00	1.00	1.00
rs6822844	3	KIAA1109	1.00	1.00	1.00	1.00
rs394581	1	RSPH3	1.00	1.00	1.00	1.00
rs548234	2	PRDM1	1.00	1.00	1.00	1.00
rs7528684	2	FCRL3	1.00	1.00	1.00	1.00
rs7528684	2	CD5L	1.00	1.00	1.00	1.00
rs3218253	2	TMPRSS6	0.15	0.28	NA	NA
rs2476601	6	HIPK1	0.13	0.58	NA	NA
rs1980422	2	RAPH1	0.44	0.69	NA	NA
rs2476601	6	MAGI3	0.49	0.98	NA	NA
rs2476601	6	RSBN1	1.00	1.00	NA	NA
rs2476601	6	DCLRE1B	1.00	1.00	NA	NA
rs2476601	6	OLFML3	1.00	1.00	NA	NA
rs2812378	2	SIGMAR1	1.00	1.00	NA	NA

Table 2.2 CD candidate genes proposed through permutation. See **Table** 2.1 for description.

Table 2.2 (continued)

Crohn's Disease						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs744166	8	STAT5B	2.0E-03	1.6E-02	2.0E-04	1.6E-03
rs744166	8	STAT5A	1.9E-04	1.5E-03	2.0E-04	1.6E-03
rs10758669	2	JAK2	2.8E-04	5.6E-04	4.0E-04	8.0E-04
rs2872507	17	CSF3	1.1E-03	1.9E-02	1.8E-03	3.0E-02
rs11465804	2	IL23R	2.0E-03	4.0E-03	2.0E-03	4.0E-03
rs11465804	2	IL12RB2	2.0E-03	4.0E-03	2.0E-03	4.0E-03
rs10045431	1	IL12B	2.7E-03	2.7E-03	2.6E-03	2.6E-03
rs2188962	8	IL5	1.5E-03	1.2E-02	3.4E-03	2.7E-02
rs2188962	8	CSF2	4.3E-03	3.4E-02	4.0E-03	3.2E-02
rs2188962	8	IL3	3.7E-03	2.9E-02	4.4E-03	3.5E-02
rs744166	8	STAT3	3.4E-03	2.7E-02	5.6E-03	4.4E-02
rs3197999	23	MST1R	3.1E-03	6.9E-02	6.6E-03	1.4E-01
rs2188962	8	IRF1	1.3E-02	9.6E-02	9.9E-03	7.7E-02
rs3197999	23	DAG1	1.1E-02	2.3E-01	1.9E-02	3.6E-01
rs2872507	17	GRB7	7.9E-03	1.3E-01	1.9E-02	2.8E-01
rs2542151	6	PTPN2	3.1E-02	1.7E-01	2.5E-02	1.4E-01
rs762421	3	AIRE	2.0E-02	5.8E-02	2.6E-02	7.6E-02
rs2066845	4	CYLD	3.1E-02	1.2E-01	3.0E-02	1.1E-01
rs2476601	6	PTPN22	2.6E-02	1.5E-01	3.9E-02	2.1E-01
rs2872507	17	IKZF3	3.9E-02	4.9E-01	4.0E-02	5.0E-01
rs10995271	1	ZNF365	4.7E-02	4.7E-02	4.5E-02	4.5E-02
rs3197999	23	TRAIP	0.07	0.81	0.07	0.79
rs2872507	17	ERBB2	0.08	0.75	0.07	0.70
rs3197999	23	BSN	0.08	0.84	0.07	0.83
rs3197999	23	IP6K1	0.11	0.92	0.10	0.90
rs3197999	23	GPX1	0.09	0.90	0.10	0.91
rs2872507	17	FBXL20	0.11	0.87	0.10	0.84
rs2542151	6	SPIRE1	0.14	0.59	0.13	0.57
rs762421	3	PFKL	0.20	0.48	0.15	0.39
rs2872507	17	PPP1R1B	0.20	0.98	0.18	0.97
rs11190140	2	ENTPD7	0.20	0.36	0.19	0.35
rs3197999	23	CELSR3	0.23	1.00	0.20	0.99
rs3197999	23	PRKAR2A	0.25	1.00	0.22	1.00
rs3828309	1	ATG16L1	0.64	0.64	0.23	0.23
rs3197999	23	CAMKV	0.27	1.00	0.26	1.00
rs11584383	3	CACNA1S	0.28	0.63	0.26	0.60
rs3197999	23	QARS	0.33	1.00	0.30	1.00
rs744166	8	TUBG1	0.35	0.97	0.30	0.94
rs2872507	17	CACNB1	0.36	1.00	0.33	1.00
rs17582416	3	CREM	0.38	0.76	0.35	0.72
rs2872507	17	PERLD1	0.39	1.00	0.36	1.00
rs11584383	3	KIF21B	0.45	0.83	0.44	0.83

Table 2.2 (continued)

Crohn's Disease						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs2476601	6	MAGI3	0.48	0.98	0.46	0.98
rs1456893	1	IKZF1	0.48	0.48	0.47	0.47
rs11747270	2	DCTN4	0.53	0.78	0.53	0.78
rs2188962	8	RAD50	0.57	1.00	0.54	1.00
rs744166	8	ATP6V0A1	0.57	1.00	0.56	1.00
rs17582416	3	CUL2	0.61	0.94	0.58	0.93
rs7746082	1	PRDM1	0.67	0.67	0.65	0.65
rs2188962	8	P4HA2	0.68	1.00	0.66	1.00
rs11584383	3	C1orf106	0.70	0.97	0.67	0.97
rs2476601	6	HIPK1	0.78	1.00	0.78	1.00
rs17582416	3	CCNY	0.79	0.99	0.79	0.99
rs11190140	2	GOT1	0.84	0.98	0.84	0.97
rs6908425	1	E2F3	0.89	0.89	0.87	0.87
rs2066845	4	SNX20	1.00	1.00	1.00	1.00
rs2066845	4	NKD1	1.00	1.00	1.00	1.00
rs2066845	4	NOD2	1.00	1.00	1.00	1.00
rs2188962	8	SLC22A5	1.00	1.00	1.00	1.00
rs2188962	8	SLC22A4	1.00	1.00	1.00	1.00
rs2542151	6	SEH1L	1.00	1.00	1.00	1.00
rs2542151	6	CEP76	1.00	1.00	1.00	1.00
rs2542151	6	PSMG2	1.00	1.00	1.00	1.00
rs2542151	6	CEP192	1.00	1.00	1.00	1.00
rs2872507	17	ORMDL3	1.00	1.00	1.00	1.00
rs2872507	17	MED1	1.00	1.00	1.00	1.00
rs2872507	17	PSMD3	1.00	1.00	1.00	1.00
rs2872507	17	RPL19	1.00	1.00	1.00	1.00
rs2872507	17	CRKRS	1.00	1.00	1.00	1.00
rs2872507	17	STARD3	1.00	1.00	1.00	1.00
rs2872507	17	NEUROD2	1.00	1.00	1.00	1.00
rs2872507	17	STAC2	1.00	1.00	1.00	1.00
rs2872507	17	TCAP	1.00	1.00	1.00	1.00
rs3197999	23	IMPDH2	0.23	1.00	1.00	1.00
rs3197999	23	GMPPB	1.00	1.00	1.00	1.00
rs3197999	23	RHOA	1.00	1.00	1.00	1.00
rs3197999	23	ARIH2	1.00	1.00	1.00	1.00
rs3197999	23	APEH	1.00	1.00	1.00	1.00
rs3197999	23	WDR6	1.00	1.00	1.00	1.00
rs3197999	23	MST1	1.00	1.00	1.00	1.00
rs3197999	23	QRICH1	1.00	1.00	1.00	1.00
rs3197999	23	IP6K2	1.00	1.00	1.00	1.00
rs3197999	23	LAMB2	1.00	1.00	1.00	1.00
rs3197999	23	USP4	1.00	1.00	1.00	1.00

Table 2.2 (continued)

Crohn's Disease						
SNP	**# Genes in Locus**	**Gene**	**$P1_{nominal}$**	**$P1_{corrected}$**	**$P2_{nominal}$**	**$P2_{corrected}$**
rs3197999	23	NDUFAF3	1.00	1.00	1.00	1.00
rs3197999	23	AMT	1.00	1.00	1.00	1.00
rs744166	8	PTRF	1.00	1.00	1.00	1.00
rs744166	8	PSMC3IP	1.00	1.00	1.00	1.00
rs744166	8	MLX	1.00	1.00	1.00	1.00
rs762421	3	DNMT3L	1.00	1.00	1.00	1.00
rs11175593	1	SLC2A13	1.00	1.00	1.00	1.00
rs11747270	2	RBM22	1.00	1.00	1.00	1.00
rs2274910	2	LY9	1.00	1.00	1.00	1.00
rs2274910	2	CD244	1.00	1.00	1.00	1.00
rs2476601	6	RSBN1	1.00	1.00	1.00	1.00
rs2476601	6	DCLRE1B	1.00	1.00	1.00	1.00
rs2476601	6	AP4B1	1.00	1.00	1.00	1.00
rs3764147	1	ENOX1	1.00	1.00	1.00	1.00
rs4263839	1	TNFSF8	1.00	1.00	1.00	1.00
rs7927894	2	C11orf30	1.00	1.00	1.00	1.00
rs7927894	2	PRKRIR	1.00	1.00	1.00	1.00
rs10758669	2	INSL6	0.55	0.80	NA	NA

The core networks involving only these candidate genes represent our mechanistic predictions of pathways underlying pathogenesis in RA and CD. From a statistical standpoint the final networks built from candidate proteins account for the excess connectivity that we initially observed: the significance remains if we restrict multigenic loci to just these genes (*direct network connectivity* $p < 2\times10^{-5}$ for RA and CD), while networks built from the remaining non-prioritized genes are less significant (*direct network connectivity* $p = 0.0368$ and $p = 0.993$, for RA and CD respectively). The remaining significance in RA is most likely a sign of additional important proteins that did not make the cutoff. From a biological standpoint, our candidates agree with experimental findings in the few cases where such evidence exists [114–121]. We therefore show that the connectivity between associated loci in RA and CD is primarily

driven by a small subset of associated proteins encoded in those regions; this observation suggests that the interacting proteins – and the biological pathways they represent – may be the targets of risk variation.

To test the biological plausibility of our nominated core networks, we asked whether the candidate genes are co-enriched in subsets of particularly relevant tissues in a reference microarray dataset consisting of 14,184 transcripts measured in 55 immune, 8 gastro-intestinal, 27 neuronal and 36 miscellaneous other tissues (126 total) [108]. These publicly available data are curated: expression intensities were converted to enrichment scores to reflect the enrichment of a gene in a tissue given its expression in all tissues. For each tissue, we compared the expression enrichment of RA and CD candidate genes to the rest of the genes in the genome using a one-tailed rank-sum test, resulting in a p-value for each tissue. A significant difference for a given tissue indicated that the genes in question were enriched for expression in that tissue compared to all genes in the genome. We also performed the same analysis for the remaining non-prioritized genes in associated regions to test whether the network prioritization method identified genes that were enriched in tissues distinct from non-prioritized genes. For discussion purposes, we defined "top" tissues as tissues achieving $p < 0.1$ (Figure 2.7 depicts the entire distribution of p-values). This analysis led to 3 main conclusions. First, we found that for each disease, enrichment only occurred in immunologically relevant tissues (Figure 2.7; strikingly, immune tissues are nearly all ranked higher than other tissues). Second, we found that this was not the case to such an extent for non-prioritized genes (Figure 2.7, black points). Third, we found that the non-prioritized genes had fewer tissues where we could detect enrichment (Figure 2.7, RA and CD candidate gene tissue scores are more

significant than tissue scores of non-prioritized genes). We formally tested this by comparing the p-value distributions for candidate genes and non-prioritized genes using a one-tailed rank-sum test ($p = 2.85\times10^{-7}$ for RA; $p = 2.55\times10^{-4}$ for CD). Of the 11 top tissues for CD candidate genes, 7 are subgroups of T-cell lymphocytes; the analogous list for RA (21 tissues) contains a mix of immune tissues, again dominated by T-cell subgroups (Table 2.3). The top tissue compartment for both diseases is defined as CD4+ T-cells.

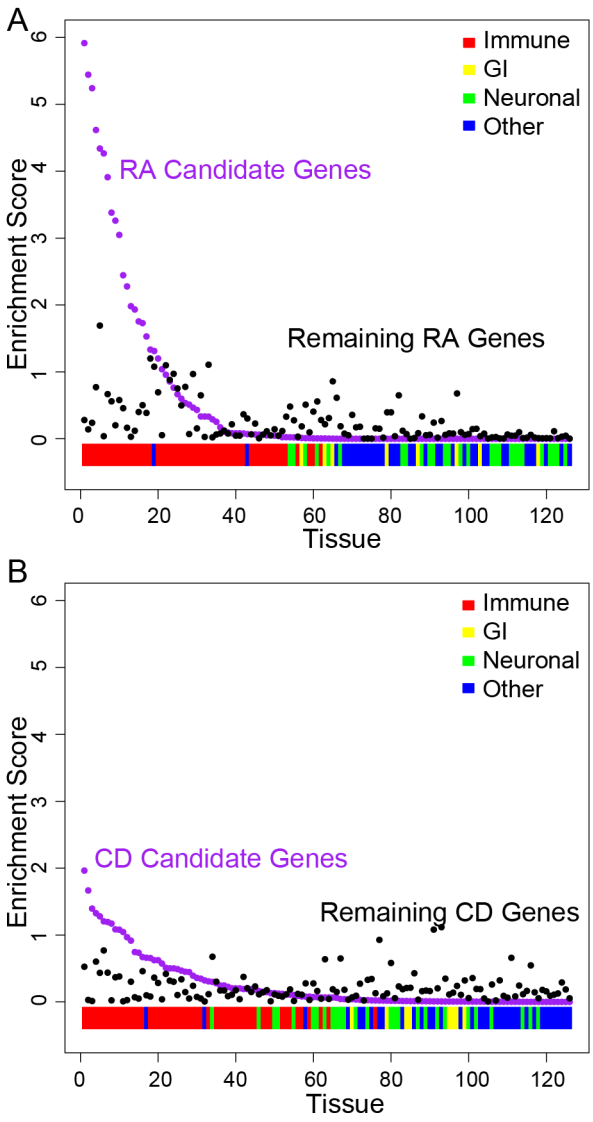


Figure 2.7 Candidate RA and CD genes are preferentially expressed in immune tissues. We obtained tissue expression data for 126 different cell types from a publicly available database, which was grouped into immune, gastrointestinal (GI), neuronal and 'other' [108]. For each tissue, we compared the expression of RA (A) and CD (B) candidate genes to the rest of the genes in the genome using a one-tailed rank-sum test, resulting in a p-value for each tissue (-log(p) is plotted on the y-axis). A significant difference for a given tissue indicated that the candidate genes were enriched for expression in that tissue compared to all genes in the genome. To test whether our

Figure 2.7 (continued)

network prioritization identified genes that were co-enriched in specific tissues beyond what was expected from all genes in associated regions, we calculated the same p-values for the rest of the genes in RA and CD associated loci (i.e., the genes that weren't prioritized via our network permutations). In this figure, we plot the tissue enrichment scores for each tissue for the candidate genes (purple) and the non-prioritized genes in the remaining regions of association (black). We indicate the category of tissue on the bottom: immune (red), GI (yellow), neuronal (green) and other (blue). We ordered the tissues by decreasing enrichment score of the candidate gene.

Table 2.3 RA and CD candidate genes are preferentially expressed in immune tissues.

Expression data was downloaded from a publically available dataset [108]. The data had been previously converted into enrichment scores (see Materials and Methods). The enrichment scores of candidate genes in RA and CD were compared to the rest of the genome by a one-tailed rank-sum test. The tissues that received a p-value of < 0.1 are shown. Of note, all tissues in this category for both RA and CD are immune, as shown in Figure 2.7.

Rheumatoid arthritis		**Crohn's disease**	
Tissue	**p-value**	**Tissue**	**p-value**
TonsilsCD4posTcells	1.21E-06	Tcellseffectormemory	0.0108
Th1	3.60E-06	TcellsBAFFpos	0.0215
TcellsCD57pos	5.74E-06	Treg	0.0402
Treg	2.42E-05	Tcellscentralmemory	0.0472
Lymphnode	4.58E-05	ThymicSPCD8posTcells	0.0525
Th2	5.42E-05	PeripheralnaiveCD4posTcells	0.0624
TcellsBAFFpos	0.000122	ThymicSPCD4posTcells	0.0637
PeripheralCD8posTcells	0.000416	MacrophageLPS4h	0.0674
Tcellscentralmemory	0.000548	MyeloidCD33pos	0.0822
Tcellseffectormemory	0.000896	PeripheralCD8posTcells	0.0831
Tonsils	0.003590	DC	0.0901
ThymicSPCD8posTcells	0.005277		

Table 2.3 (continued)

Rheumatoid arthritis		**Crohn's disease**	
Tissue	**p-value**	**Tissue**	**p-value**
ThymicSPCD4posTcells	0.010418		
PeripheralnaiveCD4posTcells	0.011665		
NKCD56pos	0.017593		
Tcellsgammadelta	0.018660		
MacrophageLPS4h	0.029525		
DC	0.046402		
Spleen	0.048697		
DCLPS48h	0.063014		
ThymicCD4posCD8posCD3pos	0.091299		

2.4.6 Crohn's Network Predicts New Loci

We hypothesized that a subset of proteins connected to the core CD network (Figure 2.8B, the network built from prioritized genes in CD loci) might be near true causal variation. Having observed significant enrichment for association in the common interactors, we then chose a more conservative approach to propose candidate genes. We selected all proteins that connect directly to the core CD network only (21 genes) and filtered them on expression in the relevant tissues (Table 2.3). While this work was being prepared, a larger meta-analysis was completed and recently published that reports 39 new loci associated to CD (295 overlapping genes) [18]. Of the 293 genes proposed by our method (small circles, Figure 2.8B), 10 were in newly associated regions (small red circles). This represents a statistically significant enrichment compared to chance expectation based on random draws from all 21,718 genes ($p = 0.001$) as well as random draws from genes expressed in at least one of the CD-relevant tissues ($p=0.01$).

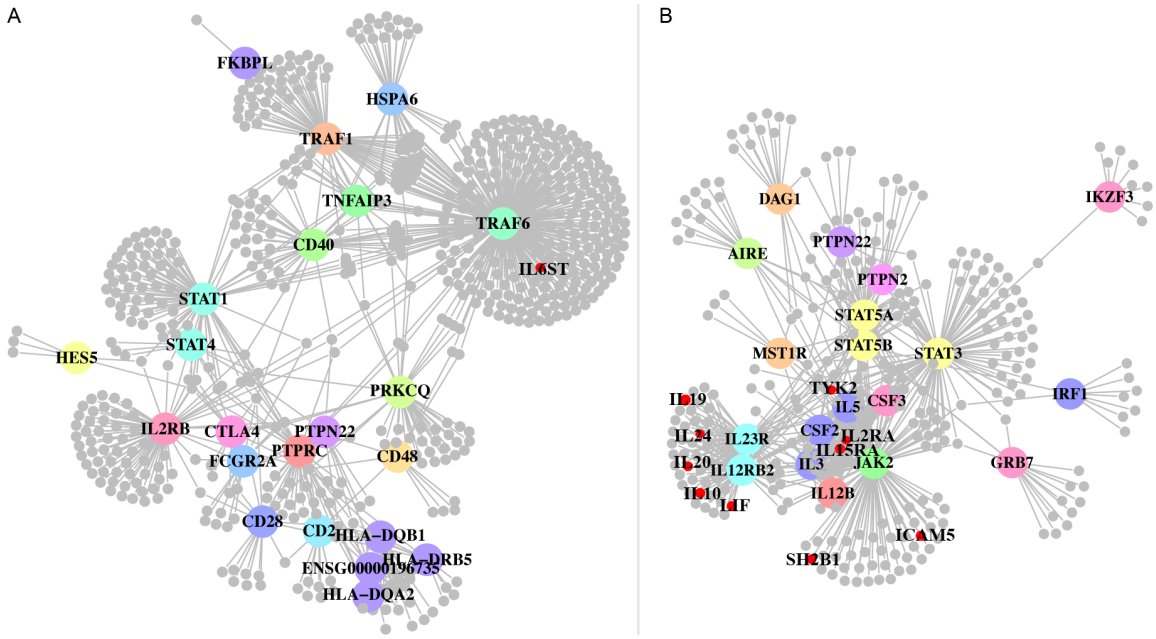


Figure 2.8 Final disease networks. Resultant networks built from candidate genes are depicted for RA and CD (**A** and **B**, respectively). Using only the candidate genes, we plotted the direct network as well as any other proteins connected to the direct network after filtering them on expression in any one of the tissues found to be specific to the core network. 610 such proteins connect to the RA network and 293 such proteins connect to the CD network. Large circles represent disease proteins, and small circles represent the connected proteins. Small red circles indicate proteins connected to the core network that were newly identified associated regions (10 proteins in CD and 1 protein in RA). The large circles are colored by locus.

We performed a similar analysis in RA since the recent meta-analysis discovered 6 new loci (18 new genes) [113]. Of the 610 genes proposed, 1 was among the 18 new genes (Figure 2.8A, small red circle). This does not represent a statistically significant enrichment.

2.4.7 Candidate gene networks suggest underlying biology

The networks (Figure 2.8) suggest pathogenic mechanisms in agreement with current thinking on disease etiology and propose novel roles for candidate proteins in these pathways. The RA network (Figure 2.10) appears to represent signaling cascades involved in the inhibition or stimulation of the NF-*k*B complex, a factor that activates transcription of genes encoding cytokines, antibodies, co-stimulatory molecules and surface receptors [119]. *STAT4* encodes a transcription factor that is activated upon engagement of cytokines, such as IL12 and interferon type I, with their receptors [119]. We show that not only does STAT4 show enrichment for connectivity, it is connected indirectly to a number of associated genes encoding surface receptor subunits that also achieve high network scores, such as IL12RB, IL2RA and PTPRC. TNFAIP3 (known as A20 in mice) is a cytoplasmic zinc finger protein that inhibits NF-kB activity, and knockout mice develop widespread and ultimately lethal inflammation, making it a plausible player in RA pathogenesis [122]. Also in the NF-kB pathway is associated protein CD40, which scores highly in our networks and binds TRAF6 and TRAF1 directly. CD40 is normally found on B cells but has also been shown to act as a co-stimulatory molecule on T cells to augment CD28 response and activate NF-kB [123].

PTPN22, a gene with strong genetic support for harboring risk variants (including the strongly associated *R620W* coding polymorphism), has been shown to act as a negative regulator of TCR but has not yet been definitively linked to a pathogenic mechanism [119,124]. Here, we place it in context of other highly associated proteins and suggest that it is part of a common mechanism.

Finally, the RA network places a number of other proteins that have not yet been formally studied in the context of the proposed network underlying RA; these include *CD2* and *CD48*, as well as *FCG2RA* and *PRKCQ*, genes suspected of being causal but not formally placed in a mechanism with other associations.

In CD the core of the candidate network (IL12B/IL23R/JAK2/STAT3; Figure 2.8B) corresponds to the interleukin-23 (IL23) signaling pathway. *IL12B* encodes p40, a component of the heterodimeric IL23. The *IL23R* gene encodes one half of the also heterodimeric IL23 receptor. This receptor is a cell surface complex found on a variety of immune cells; on activation, it induces Janus Kinase 2 (Jak2) autophosphorylation, which in turn leads to the translocation of STAT3 to the nucleus to activate transcription of various pro-inflammatory cytokines [116]. IL23 signaling is necessary for the activation and maintenance of a subset of $CD4^+$ T cells acting as 'inflammatory effectors'; these interleukin-17 responsive T-cells (Th_{17}) have been implicated in autoimmune inflammation in CD and experimental models of other autoimmune diseases [116]. We note that IL23 belongs to the interleukin 12 family of cytokines and both ligand and receptor share subunits with the canonical IL12-mediated signaling pathway, which induces activation of regulatory T cells (T_{reg}).

The CD network suggests that other proteins participate in this pathway, including the tyrosine phosphatase encoded by *PTPN2*, a gene also associated to other autoimmune diseases [125]. Other proteins that are indirectly connected to this pathway include IRF1, which we score highly and that has separately been reported to activate transcription of *IL12RB1* [126]. Furthermore, the common interactors that we prioritize for replication of association given their involvement in the CD network – including JAK1, STAT4, TYK2

and IL2RA – fall into the IL12 and IL23 signaling pathway (*TYK2* and *IL2RA* were of the genes recently found to be in regions of association).

The CD network also generates new hypotheses about potentially important genes. We prioritize AIRE, an associated protein involved in T-cell development, which has not been extensively studied in the context of Crohn's but could plausibly lead to autoimmunity. *ZNF365,* a gene that achieves a high permutation score, has been assumed to be the causal gene because it is the only gene to reside in the wingspan of its locus; however, it has not been studied as part of the core network described here (IL23R/JAK2/STAT3 pathway). Finally, *CSF2, IKZF3* and *GRB7* are in the same large locus (17 genes) but achieve significant permutation scores; these genes have been less well studied in the context of CD.

2.5 Discussion

We have shown that proteins encoded in regions associated to RA, CD, height and lipids interact and that the networks they form are significantly connected when compared to random networks. In CD and RA, the genes encoding prioritized proteins are preferentially expressed in immune tissues relevant to the pathogenesis of both diseases, while the rest of the genes in associated loci show less tissue preference. Furthermore, we can connect other associated proteins to these networks via common interactors, which appear to be encoded in genomic regions harboring further risk variants. Newly available data in CD allowed us to confirm that genes predicted to be near causal variation are indeed in regions now known to be associated to CD. We note that the conclusion of connectivity could not be extended to T2D, and we hypothesize that the lack of

connectivity may be due to disparate underlying mechanisms that have yet to be well captured genetically. Though our aim was to build and analyze networks that emerge from replicated regions of association, we feel that a promising future direction may be to look more broadly for networks enriched in weaker signals of association. Evidence that this type of analysis may be helpful is that we pointed to a set of weaker CD association signals that were found to be true positives in a larger study.

Our results have several implications for the interpretation of genome-wide association studies: first, our ability to connect the majority of associated loci in a limited number of molecular networks suggests that these represent processes underlying pathogenesis. Second, these networks are unbiased, in the sense that they do not rely on previous classifications of gene function or pathway lists; rather, we assemble our networks from low-level functional genomics data and allow network structure, if any, to emerge. Third, our approach is general; we have demonstrated it using interactions between protein products, but any relationship between genes or other genomic features (non-coding RNAs, enhancer elements, conserved regions etc.) may be used in the same fashion. Even more powerful, approaches combining such orthogonal data types will be rewarding. The limitation to using PPI data from a curated database such as InWeb is that proteins for which no high-confidence interactions exist will be left out of the analysis. As such, our analysis is limited to proteins present in the database. Additionally, while we controlled for the biases we observed, other undetected biases still may exist.

Interestingly, there are certain cases where the method is able to distinguish between proteins that are close in the genome and functionally very similar. In RA, the rs12746613 locus has 3 genes in the PPI database – *FCG2RA*, *FCGR3A* and *HSP70B*.

FCG2RA achieved a nominal p-value of 0.00703, whereas *FCGR3A* achieved p = 0.38296. Similarly, in the large rs3197999 locus in CD, the method gave *MST1R* a p-value of 0.0066 whereas *MST1*, the ligand of *MST1R*, achieved a p-value of 1. In these cases, the method is able to distinguish between functionally similar genes. There are times when it is unable to distinguish between functionally similar genes, however, such as the *IL21/IL2* locus in RA, the *STAT1/STAT4* locus in RA and the *STAT3/STAT5A/STAT5B* locus in CD.

We note in passing that the candidate genes we nominate are on average the closest to the most associated SNP in each locus, even though proximity within the LD region was not considered in the PPI analysis (p = 0.005, Figure 2.9). This supports the theory that the majority of causal variation will be close to the association signal rather than anywhere in the region of LD. We also observed overlap between genes prioritized by this method and GRAIL, a text-mining approach that uses orthogonal data [98]. We depict this information, as well as overlap between prioritized genes and the presence of non-synonymous SNPs, in Figure 2.9.

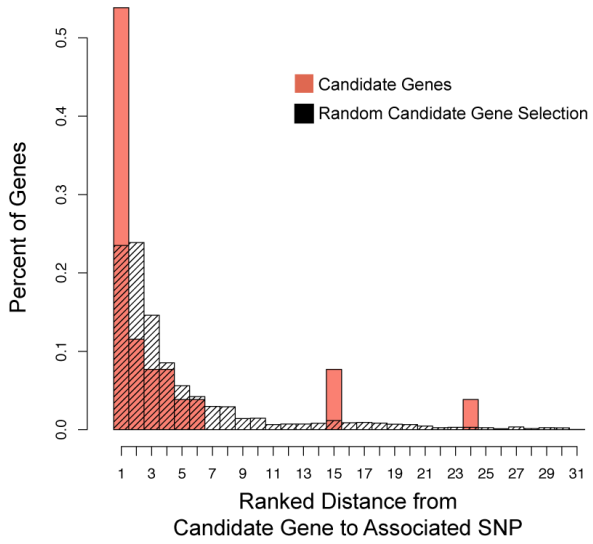


Figure 2.9 Candidate genes are likely to be near to the associated SNP. Candidate genes within multigenic loci were prioritized as described. We defined the distance from a gene to the SNP that tags it as the shortest of two distances: the distance from its start codon to the SNP and its stop codon to the SNP. Genes within a SNP's wingspan are then given ranks as to how close they are to the SNP (closest gene, 2nd closest gene, and so on). These distances were collected for RA and CD and the distribution is shown (salmon bars). We compared this distribution to the distribution of 100 simulated distances as defined by random assignment of candidate genes in associated loci (black hatched bars). The distributions are significantly different (one-tailed Kolmogorov-Smirnov test $p = 0.008$).

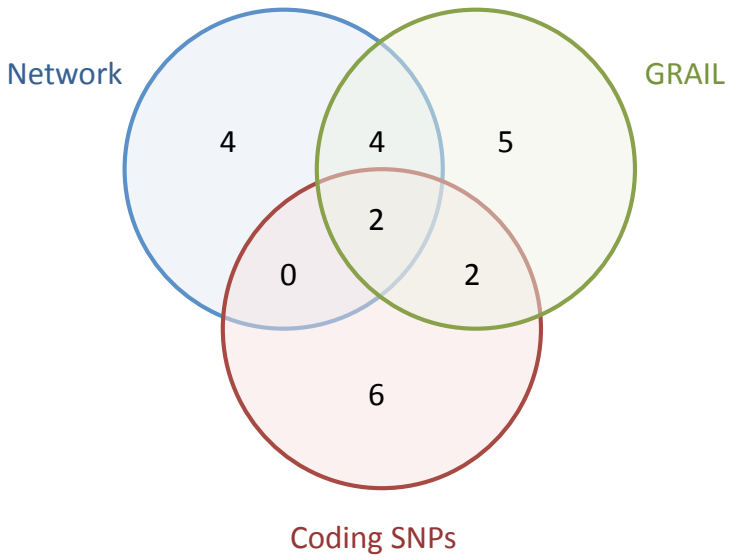


Figure 2.10 Overlap of prioritized genes across methods. For each SNP, we compared the prioritized genes through PPI networks, through GRAIL and through presence of non-synonymous SNPs. We show the overlap for all SNPs, where two methods agree if at least one prioritized gene in the region is the same.

In this chapter, we have studied 5 complex phenotypes, 4 of which show evidence of abundant PPI connections across loci. Our results therefore allow us to speculate that other complex diseases may behave in the same way and that genetic risk may be spread over the molecular processes that influence disease, rather than a single, catastrophic mutation as in Mendelian inheritance. In order to determine whether what we find here is expandable to complex disease in general, however, we would need to apply our method to the many more diseases and traits to which regions of the genome have been associated. Nonetheless, for the networks that emerge here, our approach identifies sets of proteins plausibly involved in pathogenesis, and the next step will be to identify what the molecular and phenotypic consequences of perturbing such processes are and how they relate to overall disease etiology.

2.6 Materials and Methods

2.6.1 InWeb Database

See section 1.3.5 for description. The data we used is available at www.broadinstitute.org/mpg/dapple.

2.6.2 Disease Loci

30 CD SNPs were derived from the first CD meta analysis of which 25 contain genes [16]. 28 RA SNPs were derived from the most recent RA review of which 27 contain genes [94]. 42 Height SNPs were derived from a number of analyses of which 38 contain genes [23,24,97]. 19 blood lipid level SNPs were derived from a number of analyses of which all 19 contain genes [96]. Finally, 42 T2D SNPs were derived from a number of analyses of which 37 contain genes [19–21,95].

2.6.3 Translating SNPs to genes

Hotspot and linkage disequilibrium (LD) information were downloaded from www.hapmap.org for CEU hg17 and hg18 to match the version in which associations were reported [127]. We defined the wingspan of a SNP as the region containing SNPs with $r^2>0.5$ to the associated SNP; this region is then extended to the nearest recombination hotspot. We downloaded the Ensembl human gene list from UCSC Genome Browser and collapsed isoforms into single genes [128]. We converted gene IDs from Ensemble to InWeb IDs. A gene's residence in a locus is defined by whether 110kb upstream and 40kb downstream (to include regulatory DNA) of the coding region of the gene's largest isoform overlaps the SNP wingspan[41].

2.6.4 Statistical Analysis

All analyses, including building networks and evaluating significance, were carried out in R , Perl and Python and are available at www.broadinstitute.org/mpg/dapple.

2.6.5 Author contributions and acknowledgements

Conceived and designed experiments:	Elizabeth J Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J. Xavier, Diana Tater, Chris Costapas, Mark J Daly
Performed experiments:	Elizabeth J Rossin
Contributed reagents/materials/analysis tools:	Elizabeth J Rossin, Ramnik J Xavier, Diana Tatar, Yair Benita
Optimized permutation code:	Diana Tatar, Elizabeth J Rossin
International Inflammatory Bowel Disease Genetics Consortium:	Chris Cotsapas, Mark J Daly
Wrote the paper:	Elizabeth J Rossin, Kasper Lage, Soumya Raychaudhuri, Ramnik J. Xavier, Chris Cotsapas, Mark J Daly
Contributed GWAS data:	Robert Plenge, Eli Stahl, IIBDGC
Wrote code for wingspan definitions:	Elizabeth J Rossin, Andrew Kirby
Generated manhattan plot for Figure 2.2:	Elizabeth J Rossin, Stephan Ripke
Provided additional advice and support:	David A Altshuler, Ayellet Segre, Benjamin F Voight, Josh Korn

3 Proteomic and genetic dissection of cardiac repolarization complexes

The contents of this chapter are also represented as 2 submissions to *Nature*:

Alicia Lundby*, Elizabeth J. Rossin*, Annette B. Steffensen, Christopher Newton-Cheh, Arne Pfeufer, et al. Proteomic and genetic dissection of cardiac repolarization protein complexes. *In review at Nature*.

Arking et al. Novel genetic variants influencing myocardial repolarization highlight calcium signaling. (2011) *In review at Nature.*

Individual contributions are listed at the end of the chapter.

3.1 Abstract

Myocardial repolarization is reflected in the QT interval of the heart's electric cycle, and prolongation of this interval is a risk factor for sudden cardiac death and drug-induced arrhythmias. Mendelian long QT syndrome (LQTS) is caused by rare mutations in genes important for cardiac ion channel function [129]. A recent meta-analysis of genome-wide association studies (GWAS) identified 35 loci associated with modest QT interval variation in the general population, but for most loci the causal genes remain unknown. [36,37]. Despite the biomedical implications, this reflects our rudimentary knowledge of the molecular components driving cardiac repolarization. Here, we resolve protein complexes based on five LQTS genes (*KCNQ1*, *KCNH2*, *CACNA1C*, *SNTA1*, *CAV3*) by label-free quantitative proteomics[130–132] in cardiac tissue, and integrate the complexes with GWAS data on QT interval variation from ~100,000 individuals. Twelve genes in genome-wide significant loci encode proteins in the complexes (*PLN*, *ATP1B1*, *UNC45B*, *TRAP1*, *TTN*, *CCDC141*, *ATP2A2*, *CAV1*, *CAV2*, *GOT2*, *ACTR1A*, *MYL3*; P = 1.3e-6), suggesting that these genes underlie their respective association signals. Electrophysiological recordings show that the protein encoded by *ATP1B1* modulates ion channel function, and knock-down of the orthologous gene in zebrafish shortens cardiac repolarization. Guided by proteins in the complexes, 25 single nucleotide polymorphisms (SNPs) are chosen for hypothesis-driven genotype replication in >17,500 individuals. Three SNPs that would otherwise have been missed become genome-wide significant, including one at *SRL,* a known regulator of Ca^{2+} uptake through interactions with *ATP2A2* at the protein level (rs10824026, P = 1.5e-9; rs889807, P = 1.2e-8; rs7498491, P

= 2.2e-8). Combining tissue-specific high-resolution proteomics with GWAS datasets, we show that 39% of common genetic variants associated with QT interval variation point to proteins in cardiac repolarization protein complexes. Furthermore, our approach illustrates a strategy for integrating and interpreting common and rare genetic variation using quantitative interaction proteomics.

3.2 Introduction

Modest prolongation of the QT interval duration is a quantitative heritable trait, and with the completion of the work described in the companion paper by Arking et al., GWAS have successfully identified 35 loci significantly associated with QT interval variation in the general population (≈1-4 msec/allele). In the majority of associated loci, large spans of linkage disequilibrium include many genes, and the exact genes underlying the signal remain to be identified and functionally characterized. Five of the common variant loci harbor Mendelian LQTS genes, which encode cardiac ion channels as well as proteins regulating the channel function (Figure 3.2A). Cardiac ion channels form large protein complexes and their function is greatly modulated by auxiliary subunits, in addition to the control exerted by more dynamic interaction partners, such as protein kinases[133]. However, the exact tissue-specific composition of these complexes is largely unknown. We hypothesized that some of the genetic contribution to QT interval variation could be attributed to proteins in complexes with Mendelian LQTS gene-encoded proteins (LQTS proteins hereafter), which is in accordance with results from *in silico* analyses from the companion paper by Arking et al. [manuscript in review].

To initially test the strength of this hypothesis, we ran DAPPLE (described in Chapter 2) using the published loci as well as the known LQTS genes to see whether we could identify significant connectivity based on available public data[36,37]. Briefly, we seeded the network with the 12 known Mendelian LQTS genes and seven loci harboring common QT variants (but not Mendelian genes) previously identified[134]. Consistent with the known relationships among several of the Mendelian genes, significant interconnectivity was observed (p=0.0006 for the direct connections, p=0.008 for the indirect connections, Figure 3.1).

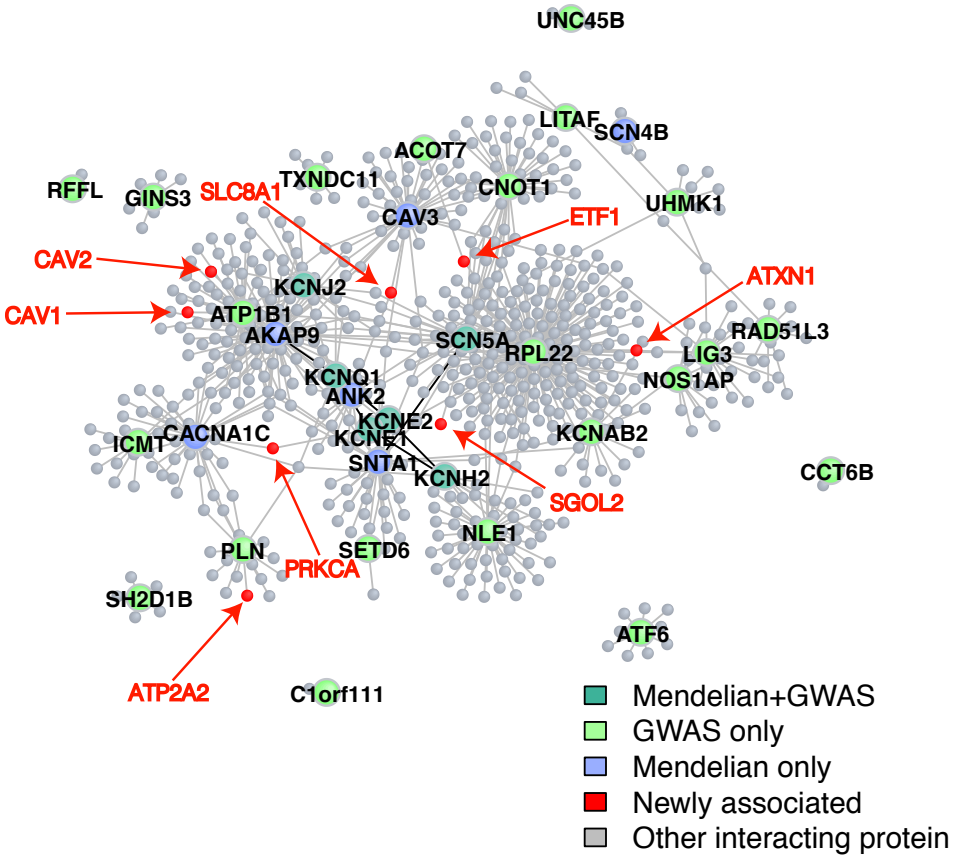


Figure 3.1 Proteins associated to QT-interval variation are significantly interconnected with Mendelian LQTS proteins and predict 8 newly associated proteins. DAPPLE was seeded with 7 previously known loci harboring common QT variants (light green), 7 Mendelian LQTS

Figure 3.1 (continued)

genes (blue) and 5 genes associated to the QT-interval both through GWAS and LQTS (dark green). Other proteins connected to the network that are newly associated in the recent QT-IGC meta-analysis are shown in red.

We identified 606 other proteins interacting directly with the seed proteins and found that 8 of them were from 7 novel loci (loci newly associated via the companion paper by Arking et al.) at genome-wide significance (*ATP2A2*, *CAV1*, *CAV2*, *PRKCA*, *SLC8A1, ATXN1, ETF1, SGOL2*; small red circles in Figure 3.1), representing significant enrichment compared to expectations under the null (hypergeometric $P = 0.03$). This likely represents an underestimate of the true biologic interactors due to the fact that several proteins in associated loci are not present in the InWeb database. Similar to the analysis described in section 2.4.3, we assigned association scores to all interacting proteins not in the novel loci and tested for enrichment in association in those genes compared to all genes in the genome from non-associated regions. We found that interacting proteins were more associated than chance expectation (rank-sum p=0.00012), suggesting that they may represent true associations yet to be discovered.

Encouraged by the *in silico* results, we went on to investigate the protein complexes associated with known LQTS proteins in an *in vivo* setting. We decided to explore a broad subset of the known LQTS proteins by affinity purification and mass spectrometry and integrated these with data on SNPs associated with QT interval variation in the general population (Digure 3.2B). We chose three ion channels and two modulators of ion channel currents as targets for the proteomic analysis. These proteins correspond to Mendelian LQTS genes both present and absent in loci identified in GWA

studies of QT interval variation. For the three ion channels chosen, mutations in *KCNQ1* and *KCNH2* cause the two most common forms of LQTS (LQT1 and LQT2), whereas mutations in *CACNA1C* cause a rare, but severe, type of LQTS associated with autism (LQT8). *KCNQ1* and *KCNH2* associated mutations result in reduced current conduction by the voltage-gated potassium channels Kv7.1[135] and Kv11.1[136], respectively, whereas *CACNA1C* mutations impede voltage-independent inactivation of the Cav1.2 L-type calcium channel[137]. The two ion channel regulators chosen, caveolin 3 (Cav3) encoded by *CAV3* (LQT9)[138] and α1-syntrophin (Snta1) encoded by *SNTA1* (LQT12)[139], have both been associated with LQTS via their influence on the Nav1.5 ion channel (encoded by *SCN5A*), but they represent a smaller proportion of LQTS overall.

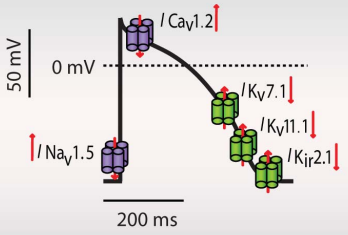
Figure 3.2 General design and experimental workflow of our integrated genetic and proteomic study. A. Left panel: Rare mutations in 12 genes can cause Mendelian LQTS. Five of the Mendelian genes reside in loci definitively associated with QT interval variation in the general population by genome-wide association studies (GWAS). Right panel: The LQTS genes encode five ion channels, Nav1.5 (*SCN5A*), Cav1.2 (*CACNA1C*), Kv7.1 (*KCNQ1*), Kv11.1 (*KCNH2*), and Kir2.1 (*KCNJ2*), mediating currents affecting the part of the cardiac action potential (black line) where the ion channels are drawn. The remaining seven LQTS genes encode proteins that interact with the illustrated ion channels and affect their currents. We resolved protein complexes of the protein products of the LQT1, LQT2, LQT8, LQT9 and LQT12 causing genes (highlighted in red). B. Top panel: We performed quantitative interaction proteomics on the five LQTS proteins to

Figure 3.2 (continued)
obtain complexes isolated from cardiac tissue. This was done by immunoprecipitating the five target LQTS proteins (baits) or control IgGs and analyzing the precipitated proteins by high-resolution tandem mass spectrometry. We identified proteins specifically and reproducibly binding to the LQTS target proteins, hereby resolving the protein complexes for each of the bait proteins. Middle panel: By integrating the proteomics data with GWAS data on common variants associated with QT interval variation we investigate the protein complexes for genetic association to this trait. Lower panel: Guided by the protein complexes we i) annotated associated loci and carried out experimental follow up on specific candidate genes and ii) identified candidate SNPs for genetic replication.

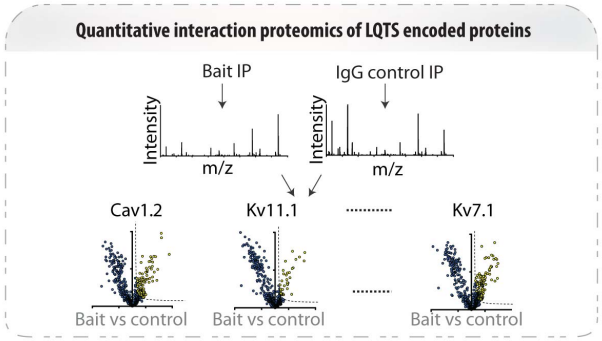
Figure 3.2 (continued)

a

Type of LQTS	Causal Gene	GWAS
LQT1	*KCNQ1*	Yes
LQT2	*KCNH2*	Yes
LQT3	*SNC5A*	Yes
LQT4	*ANK2*	No
LQT5	*KCNE1*	Yes
LQT6	*KCNE2*	No
LQT7	*KCNJ2*	Yes
LQT8	*CACNA1C*	No
LQT9	*CAV3*	No
LQT10	*SCN4B*	No
LQT11	*AKAP9*	No
LQT12	*SNTA1*	No

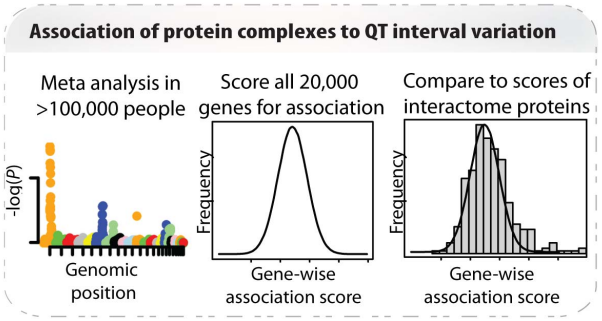


b



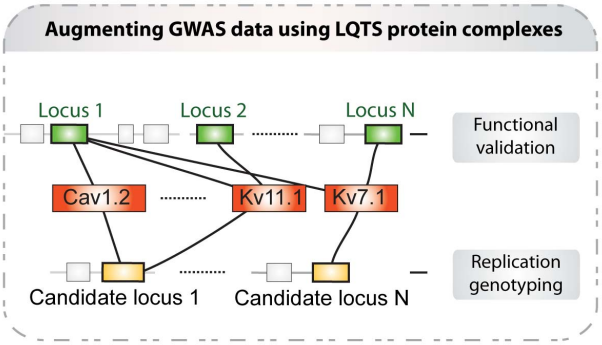
LQTS encoded proteins are immunoprecipitated from cardiac lysate and the protein composition is investigated using LC-MS/MS.

Bait IPs are compared to matched IgG control IPs to discriminate between specific and non-specific interactors.



Assign association scores to all genes in the genome.

Perform a composite test for association of the genes encoding interactome proteins using their gene-wise scores.



Red boxes show LQTS proteins; physical interactions are indicated by black lines.
Green boxes indicate genes in associated loci (QTIGC *P*<5.0e-8) that code for components in the protein complexes; these are high-priority candidates for hypothesis driven experiments.
Yellow boxes indicate genes in other loci (QT-IGC *P*>5.0e-8) that code for proteins in the complexes. These SNPs can be followed up in replication cohorts.

3.3 Results

Affinity-purification in combination with mass spectrometry is currently the most powerful and unbiased experimental method to identify the constituents of protein complexes[140,141]. We therefore applied this technique to identify interactors of the five LQTS proteins from mouse cardiac tissue. We isolated protein complexes by immunoprecipitations (IP) from total cardiac lysates generated from male mice (strain C57BL6) and studied the composition of proteins physically associated with the LQTS proteins by high-resolution and quantitative liquid chromatography tandem mass spectrometry (LC-MS/MS) using non-specific immunoglobulin G (IgG) IPs as controls. We performed triplicate IPs of all LQTS proteins (baits) and compared them to matched IgG control IPs, as quantitative mass spectrometry can effectively discriminate between specific and non-specific interactors in this experimental setting. This allowed us to identify proteins that specifically bound to the LQTS proteins by label-free quantification based on pair-wise comparison of peptide extracted ion chromatograms (XICs) and statistical significance analysis[130–132,142].

As evident from Figure 3.3A this experimental approach co-precipitates a specific cluster of proteins with each LQTS protein, and the experimental triplicates yield highly reproducible results for protein intensities (Pearson correlation coefficients $R>0.8$ in all replicates). We identified proteins specifically interacting with the LQTS proteins by a combination of the t-test derived *P*-value and the observed fold-change in LQTS-protein-to-control intensity ratios and calculated a significance curve separating specific from non-specific binders (Figure 3.3B)[130]. We defined the set of proteins surpassing a false

discovery rate cut-off at 0.05 as the bait protein complex, leading to protein complexes for each of the LQTS proteins.

As expected, the bait proteins (Kv7.1, Kv11.1, Cav1.2, Cav3 or Snta1) are among the most abundant proteins in their respective protein complexes, measured both by the XIC-based protein intensity ratios as well as by spectral counts. The protein complexes identified encompass many of the known interaction partners of the bait proteins, but more importantly they contain multiple novel components. We identify 89 specific protein interactors for Cav1.2, 31 for Kv11.1, 117 for Kv7.1, 108 for Snta1 and 334 for Cav3. These numbers are comparable to those reported for a comprehensive analysis of members of the Cav2 channel family in rodent brain, where between 97 and 161 protein interactors were consistently identified for the respective channel members[143]. We then compared the interactors identified here with those reported in the literature. For each complex, all interaction partners were determined by searching the largest protein interaction databases for interaction partners, as described previously[62,106]. Literature derived interaction partners were then compared to interaction partners identified here, using a hypergeometric distribution. In this way we determined the probability of the observed overlap given a background of ~22,000 protein coding genes. Four of five complexes are enriched for literature-derived interaction partners (Kv7.1, P= 6.0e-3; Cav1.2, P = 3.1e-5; Cav3, P = 8.9e-3; Snta1, P = 5.0e-4 using a hypergeometric distribution test)[62,106].

The consensus between our results and the literature are noteworthy given that the bait proteins are widely distributed across human tissues and that many interactors reported in the literature are identified under experimental conditions not relevant to the

biology of the cardiac ion channel complexes studied here. Moreover, as in other quantitative proteomics screens with high resolution aimed at identifying novel interactors, we do not expect perfect consensus with the literature [36,37]. For only one complex (Kv11.1) we observe no overlap with literature, which is consistent with the limited knowledge on the interactions of this protein compared to others tested in our screen.

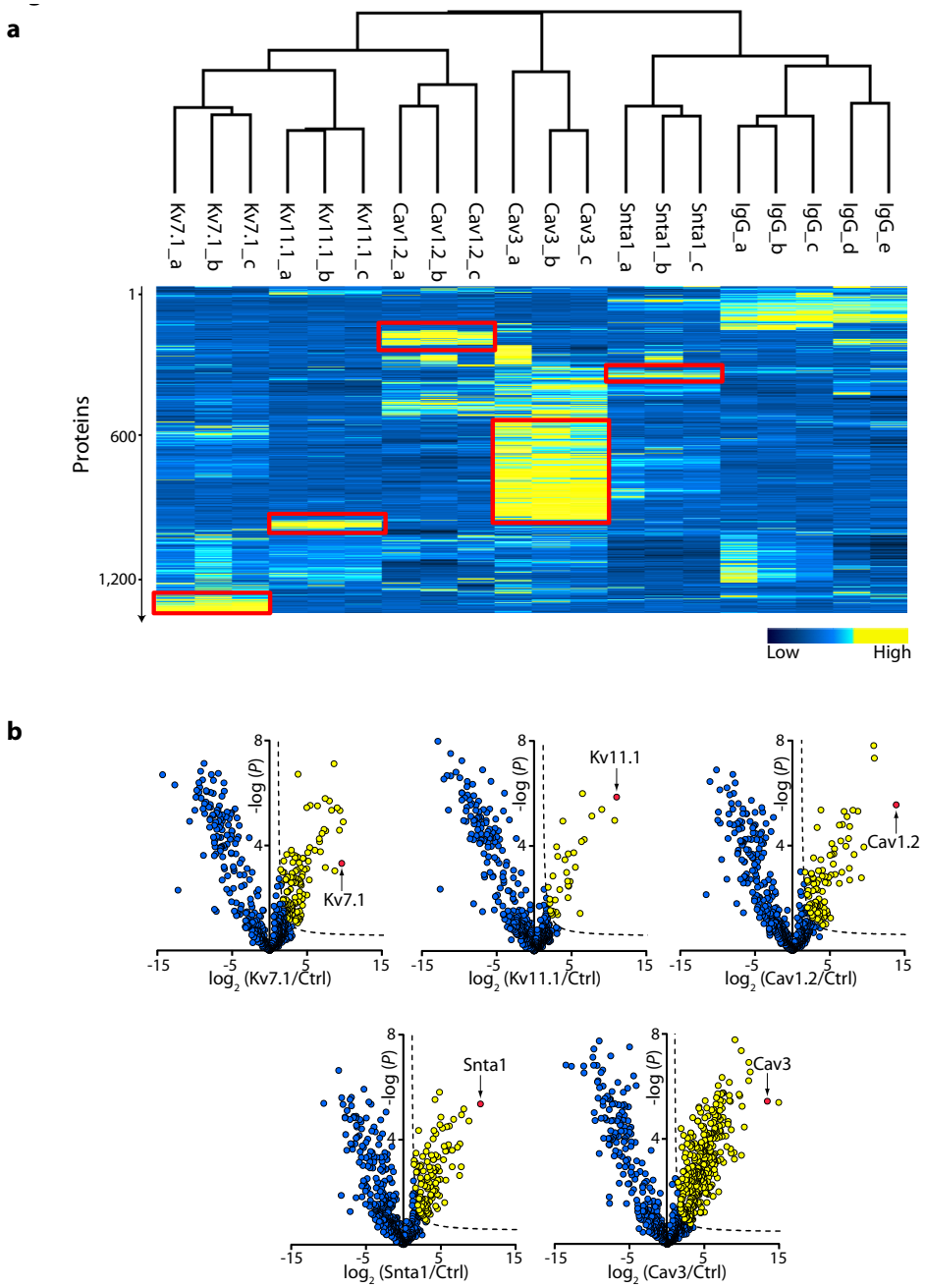


Figure 3.3 Quantitative interaction proteomics of five Mendelian LQTS proteins. A. Hierarchical cluster analysis of proteins identified in immunoprecipitation experiments visualizes the experimental specificity and reproducibility. Proteins are color-coded according to their mass-spectrometry signal intensity derived by summing the measured peptide XICs. The triplicates of the five target LQTS protein immunoprecipitations cluster together illustrating the reproducibility

Figure 3.3 (continued)
and the highlighted yellow areas indicate that each group of triplicate experiments immunoprecipitates a specific cluster of proteins. **B.** Volcano plots representing the bait versus IgG control IPs for the five target proteins. The volcano plots show negative logarithmized t-test derived *P*-values ($-\log_{10}(P)$) as function of logarithmized ratios of average protein intensities (log2) for the bait relative to control. A hyperbolic curve indicates a ratio significance *P*-value <0.05 and separates specific interacting proteins from unspecific ones. All points represent a protein with points in red being the bait proteins, points in yellow are those that specifically interact with the bait protein, and the blue points represent non-specific protein interactors. *The immunoprecipitation experiments shown in this figure were the work of Alicia Lundby.*

In chapter 2 we showed that combining protein interaction data and GWAS data can contribute to deciphering the molecular networks driving complex phenotypes and lead to identification of new candidate genes [144]. Here we exploit this method to analyze the generated protein complexes in the context of GWAS data of common SNPs underlying QT interval variation in the general population. In a companion paper by Arking et al., the QT-IGC consortium completed a GWAS meta-analysis in >100,000 individuals of European ancestry and reported 35 genome-wide significant (GWS) loci associated to QT interval variation in the general population, which together define loci spanning 154 genes. A locus is defined by identifying neighbor SNPs in linkage disequilibrium ($r^2>0.5$) to the associated SNP, and expanding to the nearest recombination hotspot as previously described [144]. Strikingly, twelve genes in the GWS loci (*PLN*, *ATP1B1*, *UNC45B*, *TRAP1*, *TTN*, *CCDC141*, *ATP2A2, CAV1, CAV2, GOT2, ACTR1A, MYL3*) encode proteins in one or several complexes. To test the joint set of proteins (737 proteins in total, 436 unique proteins) derived from all complexes for containing more GWS hits

than chance expectation, taking into account that multiple GWS proteins were represented in more than one complex, we simulated random selections of 5 complexes (each of the same number of proteins represented in the individual complexes) from all genes in the genome. After 10,000,000 random draws, we calculated an empirical p-value for the probability of selecting 22 or more GWS hits (22 represents the fact that some of the 12 GWS proteins were selected multiple times). We found the complexes to be significantly enriched for GWS genes (P = 1.3e-6). The results for the individual complexes are shown in Table 3.1. This provides a strong mechanistic link at the level of protein complexes between rare LQTS genes and a subset of genes in loci definitively associated with common QT interval variation in the general population.

Table 3.1 Enrichment in association across complexes. Each complex was tested for enrichment in proteins associated to QT-interval association according to the recent QT-IGC meta-analysis. Column 1 – bait protein with both protein nomenclatures represented. Column 2 – number of proteins pulled down. Column 3 – number of complex proteins that could be tested for association (filter out X chromosome or failed gene ID matches). Column 4 – number of complex proteins that are in genome-wide significant (GWS) loci. Column 5 – binomial test for enrichment in GWS complex proteins. Column 6 – composite test for association of complex proteins not in GWS regions.

Bait Protein	N TOTAL	N (after filtering)	N GWS	GWS Enrichment	subGWS Enrichment
CAV3 (Cav3)	358	320	11	2.8E-04	1.9E-03
CACNA1C (Cav1.2)	104	90	5	1.7E-03	2.7E-02
KCNH2 (Kv11.1)	33	30	2	3.2E-02	5.4E-02
KCNQ1 (Kv7.1)	125	116	3	9.7E-02	1.5E-01
SNTA1 (Snta1)	117	103	1	6.2E-01	2.2E-01
ALL	737	659	22	1.6E-06	1.50E-04

Because regions of the genome associated to QT interval variation are likely to code for heart-expressed genes, we considered the possibility that the association results (number of GWS proteins represented in the complexes as well as enrichment in sub-genome-wide scores) were due to enrichment for association in heart-expressed proteins rather than complex-specific proteins. Based on organ-wide proteomic mapping of phosphoproteins in rat hearts (unpublished data, Lundby et al), we collected a dataset of 2000 proteins expressed in heart tissue. We assessed the likelihood of identifying 22 GWS proteins in a random selection of 5 complexes (each of the same number of proteins represented in the individual complexes – 737 proteins in total). After 1,000,000 permutations, we found the probability of selecting >=22 heart-expressed proteins to be 0.0054, suggesting that our finding is indeed specific to complex proteins.

The ten loci contain a total of 79 genes, 12 of which the protein complexes provide experimental support for. One of these is *ATP1B1,* which is in a locus defined by rs10919070 that is convincingly associated with QT interval variation (P = 1.11e-31), and the protein product interacts with Kv11.1, Cav1.2, Kv7.1 and Cav3. Atp1b1 is well characterized as a β-subunit for the Na^+,K^+-ATPase (*ATP1A1*), but Atp1a1 was not part of any of the protein complexes, suggesting an additional unrecognized function of Atp1b1. We tested the effect of Atp1b1 on the Kv11.1 channel (encoded by *KCNH2)* by electrophysiological measurements of heterologously expressed proteins in *Xenopus laevis* oocytes. Atp1b1 affects the current mediated by the Kv11.1 channel markedly (Figure 3.4B-D). Co-expression of Atp1b1 shifts the peak of the current-voltage relationship by 10 mV to more positive potentials, slows the channel inactivation kinetics, and right-shifts the voltage-dependence of recovery from inactivation. These

data clearly show that Atp1b1 has a direct functional impact on the Kv11.1 channel properties, and suggests a biological mechanism through which common genetic variants near or in *ATP1B1* affects QT interval variation. To directly test the effect of *ATP1B1* on cardiac repolarization we used optical voltage-mapping to probe cardiac electrophysiology of *ATP1B1* zebrafish knockdown animals. Cardiac repolarization in zebrafish and humans is remarkably similar, and optical voltage-mapping enables high-resolution measurement of zebrafish cardiac repolarization [145]. Morpholino knockdown of zebrafish *ATP1B1a* (ortholog of human *ATP1B1*) results in shorter action potential duration compared to wildtype (APD80 in *ATP1B1a* knockdown 256±20 msec versus controls 321±21 msec, $P = 0.002$, see Figure 3.4E-F), further supporting *ATP1B1* as the causal gene in the rs10919070 locus.

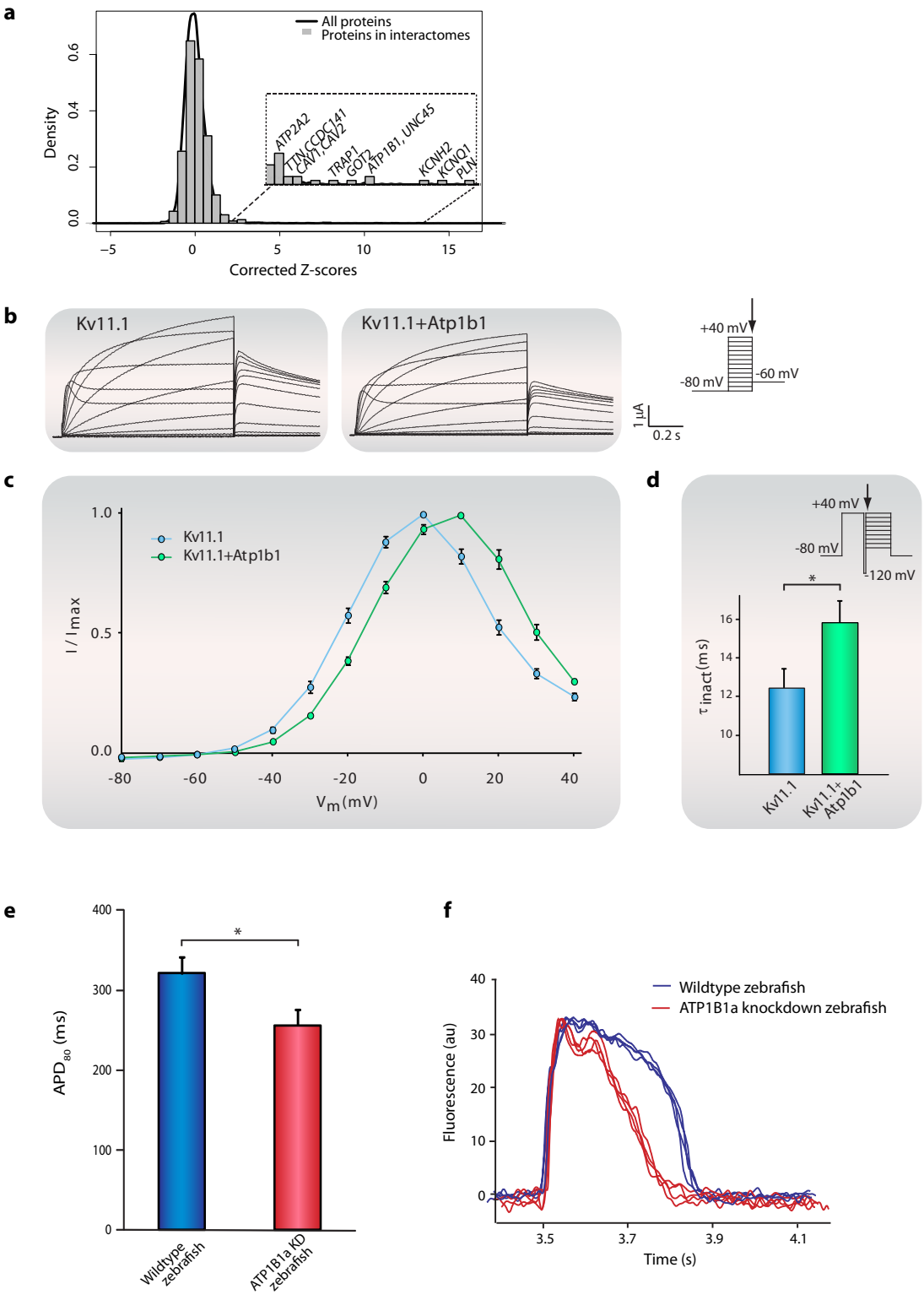


Figure 3.4 Annotation of QT-interval variation loci and electrophysiological characterization of Atp1b1-Kv11.1 interaction and *ATP1B1* zebrafish knockdowns. A. We

Figure 3.4 (continued)
used data on SNPs associated to QT interval variation from an expanded GWAS meta-analysis and replication in >100,000 individuals of European ancestry. Association Z-scores were derived for individual genes, and we depict the distribution of association Z-scores for genes represented in the complexes (grey bars) to a background distribution of all genes in the genome (black line). The x-axis represents Z-scores assigned to genes corrected for SNP density and linkage disequilibrium structure. The insert shows a zoom-in of the tail of the distribution, illustrating that the distribution is significantly enriched for genes at GWS loci (P = 1.3e-6, using permutation). **B.** One of the complex proteins in a GWS locus was Atp1b1, which has never been shown to affect ion channel currents. Representative current traces recorded from Kv11.1 (left) and Kv11.1+Atp1b1 (right) proteins heterologously expressed in *Xenopus laevis* oocytes by two-electrode voltage clamp. Step currents were elicited using the depicted voltage clamp protocol with 1s pulses to test potentials ranging from −80 to +40 mV followed by deactivation (tail) current measurements at −60 mV. **C.** Current-voltage relationships were constructed by normalizing the steady-state currents measured at the end of each voltage step to the maximum outward current and plotting it as function of the test potential. **D.** Channel inactivation kinetics was evaluated from currents elicited from the indicated pulse protocol. Inactivation time constants measured at +60 mV are shown for Kv11.1 in absence or presence of Atp1b1. Data points are mean±SEM. **E.** Morpholino knockdown of zebrafish *ATP1B1a* resulted in shortened cardiac action potentials (APD80 = 256±20 msec) compared to carrier injected controls (APD80 = 321±21 msec), n = 13 independent samples per condition. f) Superimposed normalized traces are shown for one representative sample for *ATP1B1a* knockdown (red) and control conditions (blue). * represents P<0.05. *Parts B-F in this figure were the work of Annette B. Steffensen, Alicia Lundby, Moshe Rav Acha, Stacey N. Lynch and David Milan.*

Similar to most other complex phenotypes, the currently identified common genetic variants associated with QT interval variation explain only a minority of the heritability of this trait in the population. To investigate if other proteins in the complexes could be used to guide genetic replication experiments, we excluded genes from the 35 loci definitively associated to QT interval variation and made a composite test of genetic association across the remaining genes represented in the complexes. We translated all identified mouse proteins to their orthologous human genes and derived a set of association Z-scores for each gene taking SNP density and linkage disequilibrium across and surrounding each gene into consideration[144]. Using a one-tailed Mann-Whitney rank-sum test, we compared the distribution of association scores across genes represented in the protein complexes to those for all genes in the genome. Even after excluding the 12 genes from the definitively associated loci, we find that protein complexes are significantly enriched for association (P = 1.5e-4). This suggests that subunits in the protein complexes point to genetic variants important for QT interval variation that have so far been missed. By combining genetic and proteomic evidence (association P<1e-4 or P<1e-3 and being identified as a highly abundant protein in one of the IPs), we selected 28 SNPs represented by proteins in the complexes for replication genotyping in four cohorts comprised of 17,692 independent samples in total. The proteins that formed the basis for the SNP selection are depicted in Figure 3.6A (yellow circles), along with information on which of the cardiac protein complexes they were detected in (grey lines). We selected 28 SNPs to replicate that met the following criteria: they were in LD with a gene that codes for one of the proteins pulled down in the 5 complexes, and either their association p-value was <1e-4 (23 SNPs) or it was <1e-3 as

well as the protein of interest passed a threshold for being abundantly present in one of the complexes (5 SNPs). The selected SNPs were then genotyped or looked up in four cohorts: 5,731 independent samples were genotyped in the SMART cohort, and betas, standard errors and p-values were collected for the 28 SNPs from the LifeLines cohort (n=4,865), the POSPER/PHASE cohort (n=5,135) and the RS3 cohort (n=1,961), for which the QT interval duration had been measured (in milliseconds) but the results had not been included in the QT-IGC meta-analysis. Each analysis performed a linear regression of the original QT measurement on genotype using RR-interval, age and sex as covariates. Individuals with QRS duration > 120 or history of myocardial infarction were removed. 3 SNPs were dropped due to failure in ≥ 3 of the 4 cohorts, leaving 25 SNPs that were successfully tested. The meta-analysis was done by combining betas and standard errors using the software METAL. Of those tested, 18 were directionally consistent ($P = 0.02$), 7 were nominally significant in the replication cohort ($P = 0.0003$), and 3 reached genome-wide significance when jointly analyzed with the recent QT-IGC meta-analysis (*VCL* – rs10824026, $P = 1.5e\text{-}9$; *SRL* – rs889807, $P = 1.2e\text{-}8$ and *TUFM/EIF3C/EIF3CL* – rs7498491, $P = 2.2e\text{-}8$, see Table 3.2).

SRL encodes the sarcolemmal Ca^{2+} binding protein sarcalumenin, which regulates Ca^{2+} reuptake into the sarcoplasmic reticulum by interaction with the Ca^{2+}-ATPase 2 (SERCA2)[146] encoded by the gene *ATP2A2* which itself is in a locus significantly associated to QT prolongation (rs17483, $3x10^{-12}$) [Arking et al., in submission]. The importance of *SRL* in cardiac physiology is evident from knockout mice, in which ventricular depolarization is prolonged[146]. Our data shows that SERCA2 and SRL both interact with Cav3, and that SERCA2 also interacts with the LQTS calcium channel

Cav1.2. In the accompanying paper by Arking and colleagues, 298 LQTS patients without LQT1-3 mutations were screened for mutations in 6 genes. The gene encoding SERCA2 (*ATP2A2*) was found to have stop mutations, and *SRL* was found to have amino-acid altering mutations that were likely to be damaging as predicted by two independent methods; neither *ATP2A2* nor *SRL* had damaging mutations in >300 matched controls [Arking et al., in submission]. *VCL* encodes a cytoskeletal protein, vinculin, which we show interacts with Cav3 and Snta1. Although vinculin has previously been related to dilated cardiomyopathy[147], it has never been found to be involved in QT interval variation. Our results are further supported by morpholino knockdowns in zebrafish, where we show that *VCL* knockdown has a direct affect on cardiac repolarization *in vivo* (Figure 3.5).

Table 3.2 Genetic replication results. The first three columns represent locus information of the 25 SNPs that were successfully tested for replication. Columns 4-12 represent the effect size in ms, standard error in ms and *P*-value of those SNPs in each of the QTIGC meta-analysis, in the replication cohort (17,692 samples), and in the joint QTIGC-replication meta-analysis.

Locus information			Meta-analysis			Replication			Joint		
Gene	SNP	Minor allele	Beta	SE	P-value	Beta	SE	P-value	Beta	SE	P-value
Genome-wide significant loci in the joint analysis (joint P<5e-8)											
VCL	rs10824026	A	-0.71	0.13	5.2e-8	-0.72	0.27	4.2e-3	-0.71	0.12	1.5e-9
SRL	rs889807	T	-0.51	0.10	2.6e-7	-0.53	0.22	7.2e-3	-0.51	0.09	1.2e-8
TUFM	rs7498491	A	-0.51	0.10	6.2e-7	-0.54	0.21	5.5e-3	-0.51	0.09	2.2e-8
Nominal significant loci in replication (replication P<0.05)											
CAMK2D	rs17531033	C	0.39	0.11	3.7e-4	0.66	0.24	2.8e-3	0.44	0.10	1.1e-5
TNNC1	rs352139	T	0.44	0.10	1.3e-5	0.42	0.21	2.1e-2	0.44	0.09	1.5e-6
PREP	rs7760812	A	-0.59	0.14	2.0e-5	-0.51	0.29	4.1e-2	-0.57	0.12	4.2e-6
CDH13	rs8046873	T	0.80	0.17	4.6e-6	0.75	0.45	4.9e-2	0.79	0.16	1.1e-6
Loci at P>0.05 in replication											
MB	rs17722827	A	1.00	0.20	4.4e-7	0.24	0.53	3.3e-1	0.91	0.19	1.0e-6
HSP90AA1	rs10143509	A	-0.76	0.15	5.8e-7	0.48	0.60	7.9e-1	-0.69	0.15	3.4e-6
MYO18A	rs8614	A	-0.55	0.13	2.6e-5	-0.37	0.33	1.3e-1	-0.53	0.12	1.5e-5
RPL27	rs8079855	A	0.41	0.10	8.3e-5	0.32	0.21	6.2e-2	0.39	0.09	2.6e-5
MAP4	rs777016	T	-0.46	0.11	1.2e-5	-0.13	0.22	2.8e-1	-0.40	0.09	2.9e-5
AMPD3	rs12279871	A	0.65	0.15	1.4e-5	0.17	0.31	2.9e-1	0.56	0.13	3.3e-5
DLST	rs2111705	A	0.38	0.10	5.6e-5	0.14	0.25	2.9e-1	0.35	0.09	7.4e-5
SPTBN1	rs12999048	T	-0.68	0.17	4.8e-5	-0.05	0.46	4.6e-1	-0.61	0.16	1.2e-4
PRKAR2A	rs990211	A	-0.45	0.12	1.5e-4	-0.25	0.25	1.6e-1	-0.41	0.11	1.2e-4
PABPC1	rs12114870	T	-2.00	0.48	2.7e-5	2.16	2.03	8.6e-1	-1.78	0.46	1.2e-4
ARNT	rs267734	A	-0.50	0.12	2.0e-5	0.04	0.24	5.7e-1	-0.40	0.10	1.7e-4
ALDOA	rs9924308	A	-0.38	0.10	9.4e-5	-0.11	0.20	2.8e-1	-0.33	0.09	1.7e-4
EIF3M	rs12801493	A	1.93	0.47	3.8e-5	-0.26	1.08	6.0e-1	1.58	0.43	2.3e-4
DBT/AGL	rs6682639	T	-0.79	0.23	6.8e-4	-0.76	0.54	7.9e-2	-0.78	0.21	2.3e-4
FLNB	rs6770059	A	0.68	0.17	7.6e-5	-0.03	0.44	5.3e-1	0.59	0.16	2.5e-4
PRKAR1A	rs2287301	A	0.38	0.10	1.9e-4	0.14	0.21	2.5e-1	0.33	0.09	2.7e-4
TUBA8	rs2234338	T	2.90	0.69	3.0e-5	-0.86	1.15	7.7e-1	1.89	0.59	1.4e-3
RTN4	rs6756933	T	-0.38	0.10	1.8e-4	0.71	0.28	1.0E+00	-0.25	0.10	8.9e-3

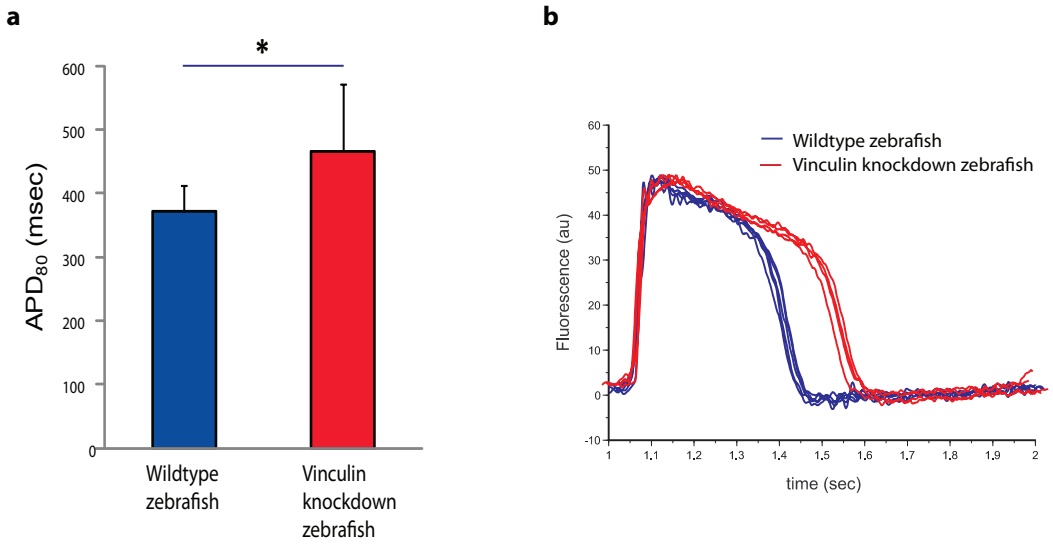


Figure 3.5 *Vinculin* knockdown prolongs action potential duration in zebrafish. A. Morpholino knockdown of zebrafish *vinculin* resulted in prolonged cardiac action potentials (APD80 =466±105 msec) compared to carrier injected controls (APD80 = 371±40 msec), *P* = 0.04, n = 13 independent samples. **B.** Superimposed exemplar traces are shown for one representative sample for *Vinculin* knockdown (red) and Control conditions (blue). *The zebrafish knock-downs were the work of Moshe Rav Acha, Stacey N. Lynch and David Milan.*

Knockdown of *TUFM* or *EIF3C* in zebrafish did not affect the action potential duration (data not shown). Based on our integrated proteomic and genetic analysis, we therefore identify three novel loci associated with QT interval variation in the general population, and for two of the loci *in vivo* evidence further supports the specific gene we prioritized as being causal. Figure 3.6B summarizes all the proteins we identified in the proteomics experiments that are encoded by genes in loci associated to QT interval variation, emphasizing the three novel loci we identified. Our data provides strong evidence that molecular interaction partners of LQTS genes contribute to common variation in the QT interval.

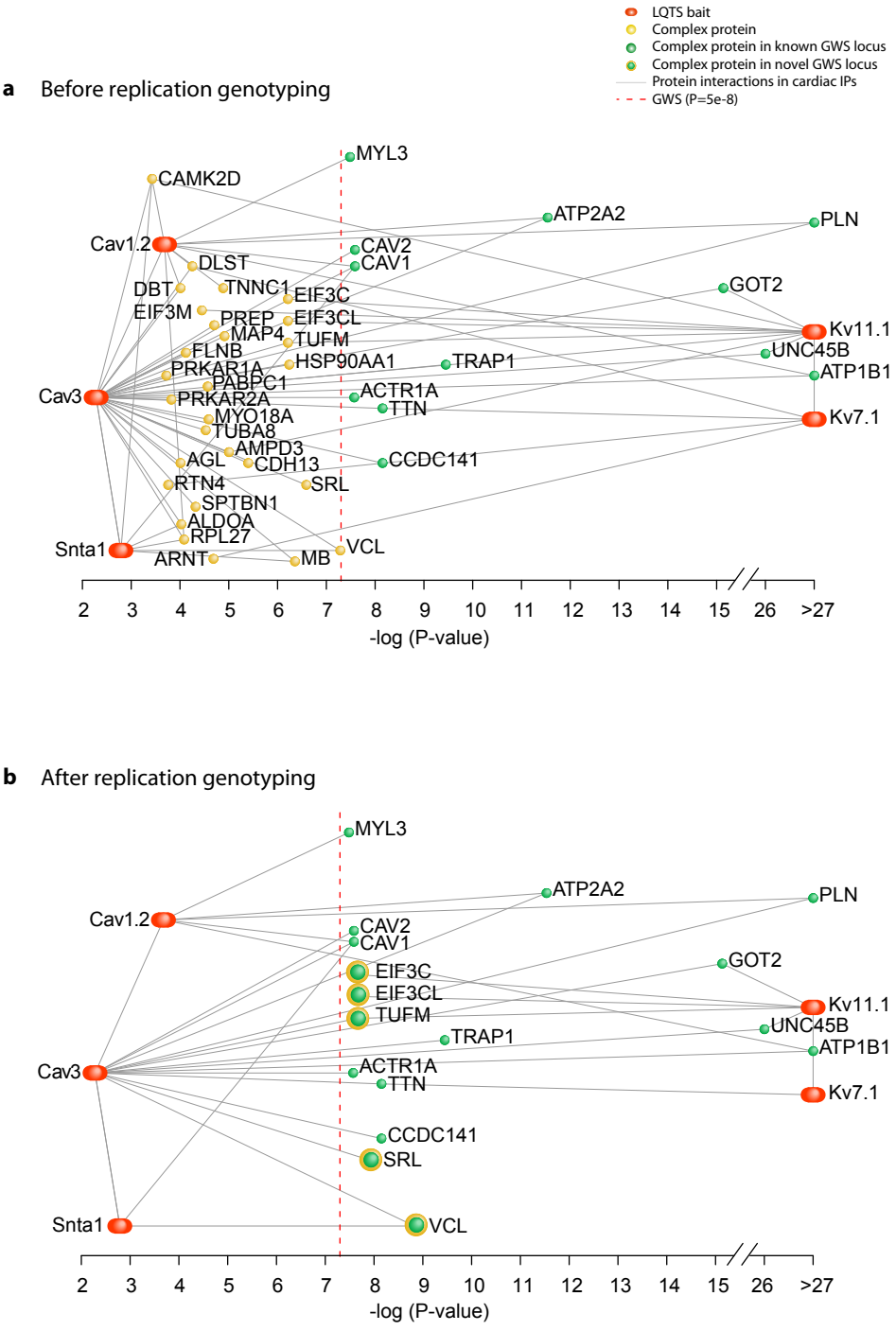


Figure 3.6 Integrative analysis of cardiac protein complexes and GWAS data. A. Depiction of the interactions identified in the proteomics experiments between the bait LQTS proteins (red) and proteins encoded by GWS genes (green) as well as proteins encoded by genes that achieved

Figure 3.6 (continued)
P<1e-4 for association to QT interval or achieved *P*<1e-3 and passed a mass-spectrometry abundance threshold (yellow). The proteins are plotted according to the best genetic association *P*-value of their corresponding genes in the horizontal direction after taking the negative 10 based logarithm of the *P*-value. Interactions are represented by grey lines. As indicated by the grey lines, several of the proteins are detected in more than one independently resolved complex. The stipulated red line indicates the threshold for GWS (corresponding to a *P*-value of 5.0e-8). SNPs representing the genes depicted in yellow were chosen for genetic replication. **B.** Complex proteins encoded by genes in GWS loci (green) highlighting the five prioritized genes from loci that became GWS after genetic replication in independent cohorts (yellow halo). Remarkably, 39% of all 38 known GWS loci (35 plus the 3 discovered here) in QT interval variation are represented by proteins that physically interact in the protein complexes resolved and illustrated here.

3.4 Discussion

Our proteomic dataset represents the first analysis of the composition of protein complexes involved in rare Mendelian LQTS based on proteins isolated from cardiac tissue. Moreover, our approach provides the first systematic overview of functional connections between genes in a Mendelian disorder and its analogous common trait using quantitative interaction proteomics. As Figure 3.6 illustrates, integrating the LQTS protein complexes with GWAS data shows that genetic variants that contribute to QT interval variation points to the interactors of proteins in which rare and highly penetrant

Mendelian variants cause LQTS. Remarkably, 15/38 (39%) of all loci now definitively associated with QT interval variation span genes that encode components of one or more of the protein complexes we dissect (Fibure 3.6B), which suggests that a very large proportion of currently

identified common genetic variation affecting QT interval duration modulates the biological systems deciphered here. Of the novel loci we discover, and associated loci we annotate, it is interesting to note that three of the genes are involved in Ca^{2+} transport (*ATP2A2, PLN, SRL*). Hereby, in addition to the well-established importance of K^+ flux, calcium signaling is highlighted as an essential component of cardiac repolarization and common QT interval variation. SERCA2 (ATP2A2) transfers Ca^{2+} from the cytosol to the lumen of the sarcoplasmic reticulum. PLN and SRL regulate the transport of Ca^{2+} into the sarcoplasmic reticulum, and the function of SERCA2 is inhibited by interactions with the former and stabilized by the latter. In addition to the proteomic and genetic evidence we provide here, *ATP2A2* and *SRL* harbor rare mutations in LQTS patients [Arking et al., in submission]. Together with our accompanying paper, the involvement of calcium signaling in cardiac repolarization is thus evidenced both from proteomics experiments, sequencing of LQTS patients, and meta-analyses of genome-wide association studies, which all converge on a cluster of physically interacting Ca^{2+} regulating proteins. The integration of several independent and orthogonal genome-scale datasets on genetics and proteomics therefore provides new insights into the molecular composition and genetic architecture of cardiac repolarization in humans. More generally, the methodological approach we have developed represents a strategy to functionally annotate loci associated with QT interval variation, for which the causal gene has not been identified, and to augment and filter modestly associated common variants. Looking forward, the methodological and statistical framework outlined here may be applicable to a number of other complex traits to elucidate their underlying biological systems and genetic determinants.

3.5 Methods

3.5.1 Tissue preparation and immunoprecipitations

6-8 weeks old male mice of strain C57BL6 were sacrificed by cervical dislocation and their hearts were harvested and snap frozen in liquid nitrogen and stored at -80 °C. Heart tissue was homogenized on a Precellys 24 and solubilized in ice-cold lysis buffer containing protease and phosphatase inhibitors. Tissue lysates were centrifuged to remove insoluble debris. For each tissue preparation produced, lysates derived from 5 mice were pooled and protein concentrations were measured by Quick Start Bradford Dye Reagent (Biorad). Solubilized heart tissue lysate was pre-cleared with Dynabeads protein G (Invitrogen) before incubation with primary antibody followed by binding to Dynabeads protein G, using either anti-Kv7.1 (SC10646, Santa Cruz), anti-Cav1.2 (AC003, Alomone), anti-Kv11.1 (AC062, Alomone), anti-Cav-3 (ab2912, Abcam), anti-Snta-1 (ab11425, Abcam) or control IgG (goat IgG: SC2028, rabbit IgG: SC2027, mouse IgG: SC2025, Santa Cruz). After washing, bound proteins were eluted with 1x sample buffer containing 100 mM dithiothreitol (70 °C, 3 min) and separated by SDS-PAGE (4-15 % Bis-Tris gels, BioRad).

3.5.2 In-gel digestion

Separated proteins were fixed in the gel (40 ml water, 50 ml acetonitrile, 10 ml acetic acid, 10 min) and visualized with colloidal Coomassie staining (Invitrogen). Each gel lane was excised and separated into four slices that were minced and destained (50 % 25 mM ammonium bicarbonate, 50 % acetonitrile) in a thermomixer (3 times 20 min, 800

rpm, room temperature (RT)). Gel dices were dehydrated (acetonitrile, 10 min, 800 rpm) followed by reduction of disulfide bonds (10 mM dithiothreitol in 25 mM ammonium bicarbonate, 45 min, RT, 800 rpm) and alkylation of cysteines (55 mM chloro-acetamide in 25 mM ammonium bicarbonate, 30 min, 24 °C in darkness, 800 rpm). After washing in 25 mM ammonium bicarbonate the gel plugs were dehydrated in acetonitrile and proteins were digested by trypsin (50 ul 12.5 ng/ul sequencing grade trypsin (Promega) in 25 mM ammonium bicarbonate for 1 hour, followed by addition of 100 ul 25 mM ammonium bicarbonate, left overnight at 37 °C). Trypsin activity was quenched by acidification of the mixture with trifluoroacetic acid to pH~2 and peptides were extracted from the gel plugs with 30 % acetonitrile in 3 % trifluoroacetic acid (30 min, 800 rpm) followed by 80 % acetonitrile in 0.5 % acetic acid (30 min, 800 rpm) and finally in 100 % acetonitrile. Organic solvents were removed by evaporation in a vacuum centrifuge. Extracted peptides were purified on STAGE-tips with two C_{18} filters.

3.5.3 Mass-spectrometry, LC-MS/MS

Peptides were eluted from the STAGE tips into 96 well microtiterplates with 2x10 ul 40 % acetonitrile in 0.5 % acetic acid and the acetonitrile was evaporated using a vacuum centrifuge reducing the sample volume to 4 ul. The peptide mixtures were acidified with 0.1 % trifluoroacetic acid in 2 % acetonitrile to an end volume of 9 ul and analyzed by on-line nanoflow LC-MS/MS. Peptide separation was performed by reversed-phase C_{18} HPLC on an Easy nLC system (Thermo Fisher Scientific) loading 5 ul samples with a constant flow of 750 nl/min onto 15 cm long analytical coloumns, packed in-house with 3 um C_{18} beads, and eluting peptides using a 135 min segmented gradient of increasing (5

%-80 %) buffer B (80 % acetonitrile in 0.5 % acetic acid) at a constant flow of 250 nl/min. The effluent from the HPLC was directly electrosprayed into an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) through a nano-spray ion source. The peptide mixture was analyzed by full-scan MS spectra (*m/z* 300-2000, resolution 30,000) in the Orbitrap analyzer after accumulation of 1,000,000 ions in the Orbitrap within a maximum fill-time of 1.000 ms with the lock mass option enabled to improve mass accuracy [131]. For every full-scan the most intense peptide ions were sequentially isolated (up to ten for every full-scan) and fragmented by higher energy collisional dissociation (HCD) in the octopole collision cell and fragments were recorded by the Orbitrap mass analyzer after accumulation of 50,000 ions with a maximum fill-time of 250 ms and using a normalized collision energy of 40%.

3.5.4 Mass spectrometry data analysis

The acquired data was processed by MaxQuant (version 1.1.1.25) (Max-Planck Institute of Biochemistry, Department of Proteomics and Signal Transduction, Munich) [132], where peptides and proteins are identified by the Andromeda search algorithm via matching of all MS and MS/MS spectra against a target/decoy-version of the mouse IPI database v. 3.68 supplemented with reversed copies of all sequences as well as frequently observed contaminants. Maximal MS/MS tolerance was 20 ppm, a maximum of 2 missed cleavages was allowed and false discovery rates were set at 0.01 both for peptides and proteins. Carbamidomethylated cysteines were set as a fixed modification, whereas N-pyroglutamine, oxidation of methionine and N-terminal acetylation were searched as variable modifications. Minimum peptide length was set at 6 amino acids. Statistical

evaluation and filtering of the resulting peptide datasets were performed in MaxQuant as previously described [132]. Protein intensities were normalized and proteins were quantified between control and case experiments by the MaxQuant label-free algorithm, resulting in LFQ (label-free quantitation) protein intensities. The downstream analysis was performed with Excel (Microsoft) and Perseus (Max-Planck Institute of Biochemistry, Department of Proteomics and Signal Transduction, Munich) software. The triplicates of each bait IP were analyzed against the five control IPs. Protein identifications were filtered for contaminants and reverse hits. A minimum of three peptide identifications with at least one being uniquely assigned to the particular protein, and protein identification in at least three immunoprecipitations were required followed by log2 transformation of the LFQ intensities. To perform statistical analysis of the label-free bait IP experiments versus control IP experiments normal distributed values were imputed for missing values using a normal distribution with width 0.3 and a downshift of the mean by 1.8 compared to distribution of all LFQ intensities. T-test based comparison of bait IPs versus control IPs were performed to identify significant interactors with P-value threshold set at 0.05 and a bend of the curve value, S0, set at 1 [133]. LFQ protein intensity ratios of bait relative to control was plotted against the negative logarithmic P-value of the t-test as was a stipulated line representing the calculated t-test based significance curve separating specific from non-specific binders. Significant interactors of the bait proteins were color coded in yellow and the rest were color coded in blue. For the hierarchical clustering, LFQ intensities were Z-scored and average linkage clustering was performed using Euclidian distance, and protein LFQ intensities were color-coded with blue representing low intensities and yellow representing high intensities. In general,

the reporting of our mass spectrometry data acquisition, processing and search results as well as sharing of all MS raw files have been done according to the *Molecular and Cellular Proteomics Guidelines*. Raw mass spectrometric files in Thermo Scientific's *.raw format are available for download through Tranche at http://proteomecommons.org using the following Hash-key:

UpjhtcVZMgE8uKwuMa6G2qQokoYYdAs2mxUAYJmrPD6HWggQ+WLr3DoMRQa
M3wyNWHjEmFyJqjIcWxioc9NVGIRub0oAAAAAAAACiA==

with password LQT1LQT2LQT8LQT9LQT12

3.5.5 Association analyses

QT-IGC: The QT-IGC consortium consists of 48 cohorts of European ancestry with QT-interval and genome-wide genotype data (>100,000 individuals in total). Each cohort contributed GWAS results from a linear regression of original QT-interval on genotype using RR-interval, age and sex as covariates (individuals with QRS-duration > 120ms or history of MI were excluded). The summary statistics (betas, standard errors and p-values) on 2.5 million SNPs (either directly genotyped or imputed) were then combined in a meta-analysis using the software MANTEL. The non-genomic-control-corrected results were used in this analysis to match what is reported in the accompanying QT-IGC study (λ_{GC}=1.069).

To test the joint set of proteins (737 proteins in total, 436 unique proteins) derived from all complexes for containing more GWS hits than chance expectation, taking into account that multiple GWS proteins were represented in more than one complex, we simulated 10,000,000 random selections of 5 complexes (each of the same number of

proteins represented in the individual complexes) from all genes in the genome. For each random selection of 737 total proteins, we counted the number of GWS hits. We then report an empirical p-value for the probability of selecting 22 or more GWS hits (22 represents the fact that some of the 12 GWS proteins were selected multiple times). To derive a p-value for each individual interacome, as described in Supplementary Figure 3, we performed a hypergeometric test, since we did not need to account for proteins being represented multiple times.

The joint test for enrichment in association performed on the remaining proteins in the complexes (those that did not achieve genome-wide significance) was carried out as described in Rossin et al. In order to control for linkage disequilibrium (LD) between genes, we broke the genome into LD blocks as defined by recombination hotspots. We then scored each block with the best association Z score achieved over that block (association data was from the QTIGC meta-analysis). This score was then corrected for the number of SNPs tested in the block using linear regression in R. The residuals from the regression were used as the corrected scores for each block, and genes were assigned scores according to the blocks they overlap. To test a group of proteins for enrichment in association, we compared the unique set of scores derived from the group of proteins to the unique set of scores for all genes in the genome using a 1-tailed rank-sum test, with the alternative hypothesis being that the group of proteins has higher association scores than scores from all genes in the genome.

3.5.6 Replication genotyping and analysis

Corhort descriptions:

SMART: The Secondary Manifestations of ARTerial disease study. SMART is a prospective cohort study among patients aged 18-74 years who are referred to the University Medical Center Utrecht, The Netherlands, because of atherosclerotic vascular disease or for treatment of atherosclerotic risk factors. The objective of the SMART study is to determine the prevalence of asymptomatic arterial disease and risk factors in patients presenting with a manifestation of arterial disease or known risk factor, and to study future cardiovascular events and their predictors in these at-risk patients. Wet-lab genotyping was carried out by KBiosciences, Hertfordshire, UK, using proprietary KASPar PCR technique.

LifeLines: LifeLines is a multi-disciplinary prospective population-based cohort study examining in a unique three-generation design the health and health-related behaviours of 165,000 persons living in the North East region of The Netherlands. It employs a broad range of investigative procedures in assessing the biomedical, socio-demographic, behavioural, physical and psychological factors which contribute to the health and disease of the general population, with a special focus on multimorbidity and complex genetics.

PROSPER/PHASE: All data come from the PROspective Study of Pravastatin in the Elderly at Risk (PROSPER). A detailed description of the study has been published elsewhere. PROSPER was a prospective multicenter randomized placebo-controlled trial to assess whether treatment with pravastatin diminishes the risk of major vascular events in elderly. Between December 1997 and May 1999, we screened and enrolled subjects in Scotland (Glasgow), Ireland (Cork), and the Netherlands (Leiden). Men and women aged 70-82 years were recruited if they had pre-existing vascular disease or increased risk of

such disease because of smoking, hypertension, or diabetes. A total number of 5,804 subjects were randomly assigned to pravastatin or placebo. A large number of prospective tests were performed including Biobank tests and cognitive function measurements. Resting 12 lead ECGs were recorded at baseline and annually thereafter and were analyzed using the University of Glasgow analysis program. A whole genome wide screening has been performed in the sequential PHASE project with the use of the Illumina 660K beadchip. Of 5,763 subjects DNA was available for genotyping. Genotyping was performed with the Illumina 660K beadchip, after QC (call rate <95%) 5,244 subjects and 557,192 SNPs were left for analysis. These SNPs were imputed to 2.5 million SNPs based on the HAPMAP built 36 with MACH imputation software.

RS3: The Rotterdam Study III (RS-III) is a prospective population-based cohort study. The cohort comprises 3,932 subjects aged 45 years and older, living in the Ommoord district in Rotterdam, the Netherlands. The Medical Ethics Committee of Erasmus Medical Center approved the study and written consent was obtained from all participants. Electrocardiograms were recorder on ACTA electrocardiographs (ESAOTE, Florence, Italy) and digital measurements of the QRS intervals were made using the Modular ECG Analysis System (MEANS). All RS-III participants with available DNA were genotyped using Illumina Human 610 Quad array at the Department of Internal Medicine, Erasmus Medical Center following manufacturer's protocols. Participants with call rate < 97.5%, excess autosomal heterozygosity, sex mismatch, or outlying identity-by-state clustering estimates were excluded. After quality control 2,082 RS-III participants were included. Of these, 1961 participants were included in this study.

For the SMART data, we ran a linear regression in Plink to test for association to the duration of the QT interval in the same manner as was done in the QT-IGC meta-analysis as well as the other 3 cohorts, controlling for age, sex and heart rate and excluding individuals with QRS duration > 120 or past history of MI. Association results are expressed in terms of a 1-tailed p-value in the replication cohort and a 2-tailed p-value when folded in with the meta-analysis. The meta analysis was done with the program Metal using effect size estimates and standard errors. These results are reported in the main text as Table 3.2. We assessed the results as follows: first, we counted the number of SNPs that were nominally significant ($P < 0.05$) in the replication cohort. 7 were nominally significant. 1.25 SNPs by chance are expected to be nominally significant, and this therefore represents an enrichment at P=0.0003 using a binomial test. We then did a sign-test for directional consistency, and found that the effect sizes of 18/25 SNPs were directionally consistent with QTIGC ($P = 0.02$). Then, we considered the replication p-value in addition to direction of effect by counting the number of SNPs that improved the QT-IGC meta-analysis p-value when jointly considered. 11 improved the original QT-IGC p-value, whereas on average 7.6 are expected by chance based on simulation ($P = 0.03$). Finally, three novel SNPs reached genome-wide significance when folded in with the QT-IGC meta-analysis: rs10824026, P = 1.5e-9; rs889807, P = 1.2e-8; rs7498491, P = 2.2e-8.

3.5.7 Electrophysiology and data analysis:

Preparation and injection of cRNA into *Xenopus* oocytes, purchased from EcoCyte Bioscience (Castrop-Rauxel, Germany), were done as described [18]. cDNAs were verified

by sequencing. GeneBank accession numbers of the clones used were NM_000238 for hKv11.1a and NM_001677 for hATP1B1. Currents were recorded from three batches of oocytes injected with hKv11.1a, hKv11.1a+hATP1B1 or hATP1B1 cRNA with hKv11.1a and hATP1B1 injected at a 1:1 molar ratio from a holding potential of −80 mV. Electrophysiological recordings were performed at room temperature (22°C–24°C) 3 days after injection in Kulori medium (90 mM NaCl, 4 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 5 mM HEPES, pH 7.4) using a two-electrode voltage clamp amplifier (CA-1B, Dagan, Minneapolis, MN, USA). Data analysis was performed using Pulse (HEKA, Lambrecht, Germany), Igor Pro 4.04 (Wavemetrics, Lake Oswego, OR, USA), and GraphPad Prism (GraphPad Software Inc, San Diego, CA, USA). All values are displayed as mean ± SEM. Current–voltage (I/V) relations were obtained from the step-protocol by plotting the outward current at the end of the second test-pulse as a function of the test-potential. Inactivation kinetics was evaluated by the time constant derived from a monoexponential fit to the decaying phase of the current. The voltage-dependence of activation, inactivation and recovery from inactivation was determined by fitting normalized currents versus test potentials to a two-state Boltzmann distribution of the form $I(V) = 1/(1+\exp[(V_{½} - V)/a])$, where $V_{½}$ is the potential for half-maximal activation and a is the slope factor. The number of independent experiments is indicated by n. Comparison of the biophysical properties in the presence and absence of hATP1B1 were performed using an unpaired t-test with $P < 0.05$ being considered significant.

3.5.8 Zebrafish experiments:

TuAB or Ekwill wild type zebrafish strains were reared according to standard techniques. At the single cell stage, fertilized oocytes were injected with 1-10ng of antisense morpholino oligos targeting the transcription initiation sites of ATP1B1a [19], vinculin [20], TUFM (5' - GAATTTTATAACTTACCGGAGAGGC – 3') or EIF3C (5' – GTCTTCTCCACAAACTCACTGCTGT – 3') dissolved in Danieau's solution (58 mM NaCl, 0.7mM KCl, 0.4 mM MgSO4, 0.6 mM Ca(NO3)2, 5.0 mM HEPES pH 7.6). Controls were injected with Danieau's solution alone. Embryo hearts were microdissected, stained with di-4-ANEPPS (Invitrogen) and imaged on a CCD Camera (Cardio-SMQ, Red Shirt Imaging) at 1000 frames per second as previously described [21]. Cardiac motion was arrested with the use of 15uM blebbistatin (Sigma), field pacing was employed to control beating frequency (Grass S48 Stimulator).

3.6 Acknowledgements

The work presented in this chapter is the result of a large collaboration across multiple institutions.

Overall idea, concept, and project coordination:	Alicia Lundby, Elizabeth J Rossin, Kasper Lage, Jesper Van Olsen
Conceived and designed the immunoprecipitations and proteomics experiments:	Alicia Lundby, Elizabeth J Rossin, Kasper Lage, Jesper Van Olsen
Performed the immunoprecipitations and proteomics experiments:	Alicia Lundby
Analyzed the proteomics data:	Alicia Lundby and Jesper Van Olsen
Conceived and designed electrophysiological experiments:	Alicia Lundby
Performed and analyzed the electrophysiological experiments:	Annette B. Steffensen
Conceived and designed statistical enrichment analyses:	Elizabeth J Rossin, Kasper Lage
Performed enrichment analyses:	Elizabeth J Rossin
Identified SNPs for replication:	Elizabeth J Rossin, Alicia Lundby, Paul I. de Bakker, Kasper Lage, Jesper Van Olsen
Conceived and designed genetic replication experiments:	Elizabeth J Rossin, Mark J Daly, Paul I. de Bakker
Performed genetic meta-analysis:	Elizabeth J. Rossin.
Conceived and designed zebrafish experiments:	David Milan, Patrick Ellinor
Performed and analyzed the zebrafish experiments:	Moshe Rav Acha, Stacey N. Lynch.
Contributed meta-analysis GWAS data:	QT-IGC, Chris Netwon-Cheh, Paul I. de Bakker, Daniel Arking, Sara Pulit.
Contributed input for the manuscript:	Soren Brunak, Soren-Peter Olsen, Chris Netwon-Cheh, Pim van der Harst, Paul I. de Bakker.

Contributed with data for genetic replication:	SMART: Folkert W. Asselbergs, Paul I. de Bakker LifeLines: Piim van der Harst PROSPER-PHASE: J. Wouter Jukema, Stella Tromet, Ian Ford, Peter W. Macfarlane RS3: Bouwe Krijthe, Albert Hofman, Andre Uitterlinden
Wrote the paper	Elizabeth J Rossin, Alicia Lundby, Kasper Lage, Jesper Van Olsen.

4 Integration of protein-protein interaction data with rare variation

An abbreviated version of section 4.3.1 of this chapter also appears in *Nature* as:

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* Available at: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature11011.html. Accessed 5 April 2012.

Individual contributions are listed at the end of the chapter.

4.1 Abstract

Owing to recent technological advances, genome sequencing in medical genetics to discover disease-relevant variants is now a reality. Though their contribution to disease architecture remains unclear, investigators are now beginning to study the role of rare single nucleotide polymorphisms (SNPs) in medical genetics. Not surprisingly, a number of recent publications have shown that significant analytic challenges exist when trying to detect association to such rare events. Recent studies showed that the problem can be addressed by testing groups of variants together, though each acknowledges that the appropriate functional unit over which to group variants is still not clear [148,149]. Since genes are well-mapped (at least in comparison to other functional units in the genome), grouping rare variants over a gene could be very powerful, but it has become apparent that most methods still cannot overcome the burden of multiple testing, even in the setting of testing SNPs within genes jointly[150]. We showed in chapters 2 and 3 that common variants associated to complex traits tend to be near genes that relate to one another through protein-protein interactions. We hypothesize that rare variants contributing to risk for disease likely affect modules of proteins in a similar way and that studying rare genetic variation at the level of biological pathways (as opposed to individual genes or variants) through joint analysis of sequencing data with protein-protein interaction (PPI) data will reveal insight into causal genes and networks affected by rare variation. Here, we investigate the role of PPI data in analyzing rare variation. First, we show that genes harboring functional *de novo* variants in autistic patients form PPI networks by leveraging DAPPLE, a method discussed in chapter 2. Second, we extend this concept to case/control exome sequencing, where we simulate networks

affected by rare variation and apply a data-biased random-walks method (herein called "DAPPLE/SEQ") that augments the CALPHA gene-burden test by jointly considering PPI and genetic data. In both settings, we provide a principled approach to show that PPI has promise in highlighting important genes and networks affected by rare variation that may have been otherwise missed in variant testing alone.

4.2 Introduction

Advances in high-throughput sequencing have now reached a point where it is economically feasible and technically possible to study rare (MAF < 5%) and *de novo* genetic variation in thousands of patients. For diseases that are highly polygenic, this task poses a significant challenge. It is likely the rare alleles contributing to complex traits are of modest effect, as evidenced by the lack of results coming out of linkage studies[2]. Unlike common variation, rare alleles of low effect are often too individually infrequent for a traditional test of association to be well powered enough to detect association or to distinguish it from the many neutral variants nearby. Even more challenging is the analysis of *de novo* variation, which is rarely seen more than once in a single cohort in the same gene, let alone at the same position in the genome[151].

One of the most important areas for development is how to analyze and interpret the incredible amount of genetic data that is emerging. The rarity of individual causal single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (indels) and the number of genes likely to be contributing to risk for complex disease compounded with the sheer number of rare variants in the human population makes it difficult to distinguish causal from neutral variants. For example, results from on-going

autism case/control sequencing studies in our lab show that 95% of all genes contain at least one rare non-synonymous SNP in 789 people [unpublished data]. To address this challenge, a number of burden tests have been developed over the past few years to overcome the power limitations in analysis of rare variation. These approaches can be categorized into three groups: collapsing markers across a region, weighting markers based on functional categories and distribution-based approaches[150]. Li and Leal proposed collapsing variants over a region, and Madsen and Browning augmented this approach using a weighted-sum statistic, whereby markers are weighted according to allele frequency, and the weights are summed over genes[149,152]. Other methods use functional measures such as alteration in protein function to weight alleles[153]. More recent methods use changes in distributions of rare variants as the test of association, such as the binomial over-dispersion test called CALPHA for dichotomous traits or SKAT for continuous traits[148,154]. Yet, results from GWAS have consistently revealed that there are likely hundreds of causal genes underlying complex traits, which would predict that only a small fraction of cases harbor rare mutations in any one of them. This may explain why a recent study found these methods to be somewhat underpowered in the context of simulated and real genetic data[150]. In this case, it is clear that more innovative methods for interpreting genome sequence data are required.

We hypothesize that for complex traits, causal genetic variation affects a common but limited set of underlying molecular processes that modulate risk to disease. We therefore suggest that it will be necessary to consider rare variants in the context of functionally connected genes to detect association. Autism, a pervasive developmental disorder characterized by language delay, restricted interests/repetitive behaviors and

social impairment, is an example where GWAS has as of yet not identified common variation contributing to disease; one naturally emerging hypothesis is that the genetic architecture of autism may include rare variation. Recent pathway studies revealed that rare CNVs in autism affect genes that share functional relatedness beyond chance expectation; another study detected abundant connections between established autism proteins via yeast-two-hybrid assays[55,155]. Likewise, we and others have shown that loci associated to other complex traits (such as Crohn's disease, rheumatoid arthritis, blood lipid levels and QT-interval variation) include genes that form significant physical interaction networks or fall into similar pathways [87]. We propose that this concept be applied more generally to ongoing sequencing efforts in complex traits. To our knowledge, there has been little rigorous testing of the use of PPIs in analyzing rare and *de novo* variation revealed by exome sequencing.

In our joint analysis of rare variation and PPI data, we will use the publically available PPI database InWeb (described in chapter 2) consisting of 169,810 interactions across 12,793 proteins. Investigators have rapidly populated databases of such PPIs over the past decade, mainly through manual curation of the literature[60,61,64–66,80]. The data is noisy – beyond technical false positives, *in vitro* evidence of binding may not recapitulate *in vivo* binding due to temporal and spatial expression differences[105]. InWeb is a meta-database that addresses many of the concerns about false positives in PPI data by assigning a confidence score to each interaction (see chapter 2) [62,106].

This chapter will consider two types of rare variation: *de novo* variants discovered through analysis of trios and rare inherited variants discovered through case/control sequencing. The challenge in interpreting *de novo* variation is that unless the relevant

genes are affected in multiple people within a sequencing cohort, it can be difficult to pinpoint which genes to follow up on, especially if *de novo* variation is predicted to play a relatively small role in disease etiology. To consider the utility of PPIs in the context of *de novo* variation in autism, we will describe the application of a tailored version of DAPPLE to a recent autism *de novo* discovery sequencing effort. Autism is one of the most heritable complex disorders, suggesting that genetics plays a leading role in its etiology. Though *de novo* variation cannot contribute to this heritability, it is likely that causal *de novo* variants affect similar pathways as inherited ones.

The other category of variation discussed here is rare inherited variation discovered through case/control exome sequencing which may contribute to heritability. Whereas *de novo* variation can be easily fitted to the previously benchmarked method DAPPLE, analysis of case/control rare variant data will require significant methodological development. The goal of such an analysis is to find sub-networks in the InWeb database that are enriched for association to rare variation. Ideally, we would test all possible sub-networks; however, this naïve solution not only would require a astronomical multiple testing burden to overcome but also has been shown to be NP-hard[156]. Methods to discover "enriched sub-networks" – groups of proteins enriched for both connectivity and an independent metric – have been studied over the past decade. Promising approaches include simulated annealing, an adaptation to the Prize-Collecting Steiner Tree problem and data-biased random walks (DBRW) [156–158]. Compared to the former two methods, DBRW is an attractive approach due mainly to its technical feasibility.

Komurov et al. have previously shown that DBRW is able to recover enriched sub-networks when integrating genetic and gene expression data[158]. The method approaches the problem by transforming the PPI database into a transition matrix, whereby the transition probability from one node to the next for a random walker is dependent on the association of each node[148]. Clusters of interconnected nodes that are enriched for the independent score should be visited more often than other nodes. Inspired by the success of Komurov et al.'s approach and the practicality of its fast implementation, here we adapt this approach to analysis of rare variation associated to disease.

4.3 Results

4.3.1 *De novo* variation in autism and the protein complexes implicated

To identify the role of *de novo* exonic point mutations in autism, our group sequenced whole exomes of 96 trios (affected offspring and parents) from mostly simplex families, followed by a second wave of 77 trios. Neale and colleagues generated a QC and analysis pipeline that resulted in 93% of the exome being assayable due to sufficient depth of coverage and a 98% validation rate on variants called as *de novo*. Across both waves, they identified 163 coding single nucleotide variants consisting of 101 missense, 10 nonsense, 2 conserved splice-site and 50 silent mutations. An exome mutation rate of 1.5e-8 per base per person was estimated, which does not represent a significantly higher mutation rate in autistic individuals than expected by chance.

While no significant evidence that *de novo* events play a major mechanistic role in this sample was observed, it does not preclude the variants from conferring significant

autism risk. Recent CNV studies highlight the inescapable fact that hundreds of loci are involved in autism and some severe *de novo* mutations affecting these genes likely contribute to autism risk in specific individuals[55]. Therefore, more modest contributions of *de novo* variants, such as 10-20% of cases carrying a risk-conferring event, could be consistent with the observed data. That is, while the set of data are not significantly different from chance mutation, the subset of genes that harbor *de novo* mutations could be enriched for autism relevant genes.

Of the group of genes harboring *de novo* functional (non-silent) mutations (113 in total), we asked whether the protein products are unusually connected to each other using DAPPLE. The consideration of *de novo* variation introduces a new bias that the current permutation methodology (described in Chapter 2) does not account for: large mutation target size. *De novo* variants are more likely to be observed in larger genes, since more coding real estate offers more chances for a *de novo* mutation. Therefore, we set out to test whether the PPI data is non-randomly distributed according to gene size. First, we calculated the correlation between gene size (excluding non-exon regions) and the respective protein binding degree in the InWeb database. A significant but very small correlation exists (Figure 4.1, $R^2 = 0.00148$, $p = 4.12e\text{-}5$). Importantly, DAPPLE will correct for this type of bias, since it will compare to networks of the same binding degree distribution.

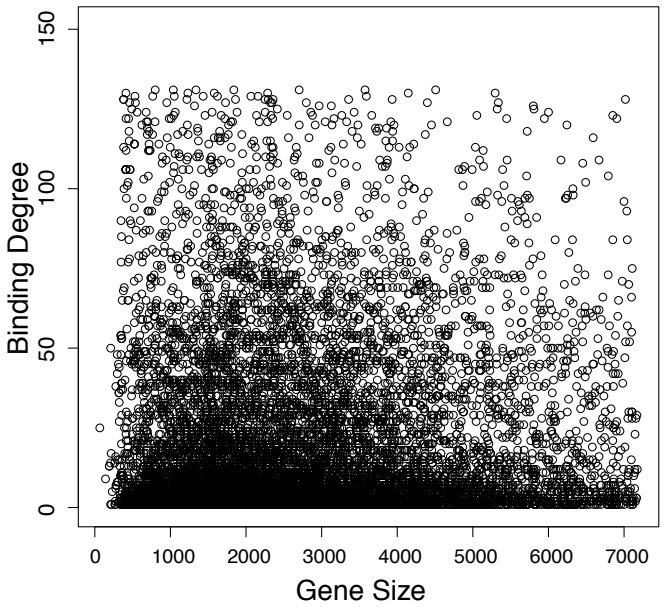


Figure 4.1 Correlation between gene size and binding degree. To test whether large genes are likely to have more binding partners, we calculated the Pearson correlation between gene size (defined by exons in kilobases) and binding degree in the InWeb database. We found a small but significant correlation ($R^2 = 0.00148$, $p = 4.12e\text{-}5$).

We then considered the possibility that a larger correlation structure exists between the size of a gene and the size of its binding partners (ie, big genes bind big genes). Interestingly, we found a larger correlation (Figure 4.2, $R^2 = 0.03$, $p < 2e\text{-}16$), suggesting that the size of a gene is correlated with the size of its binding partners. This means that the DAPPLE p-values are confounded (albeit very modestly) by gene size *if* there is a systematic bias in the size of the input genes. This is typically not the case for GWAS but will be the case for *de novo* variation and CNV studies, since they will tend to be biased toward larger genes.

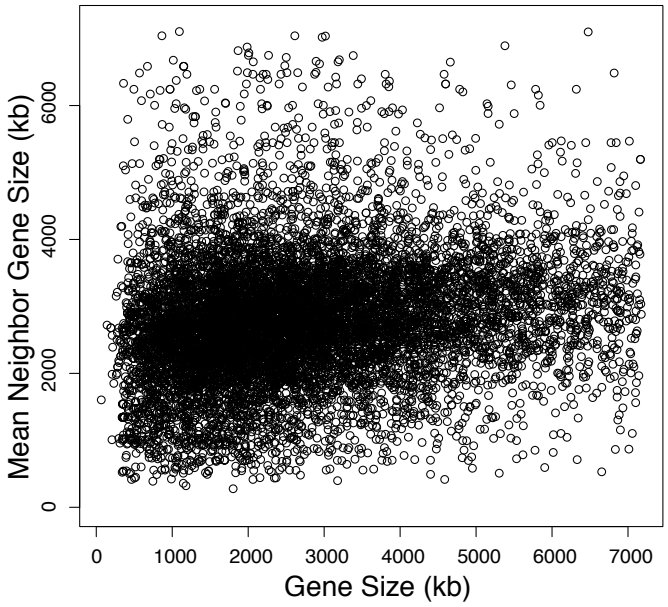


Figure 4.2 Correlation between gene size and mean neighbor gene size. We found a modest and significant correlation between a node's gene size (defined by exons in kilobases) and the mean of its neighbor's sizes (Pearson R^2=0.0354, p<2e-16).

To correct for this bias empirically, we generated a list of random *de novo* point mutations using a mutation rate model that used fixed differences between humans, chimps and baboons to estimate the relative frequency of all possible single-nucleotide changes as a function of the bases to the left and right (i.e., all 64x3 possible three base changes XY_1Z -> XY_2Z) and sampled sets of 113[159,160]. The final p-value assigned to a network becomes the empirical p-value based on 1000 simulations representing the frequency with which an equal or better DAPPLE direct network significance score (each based on 10,000 DAPPLE permutations) was observed.

We found 22 direct connections amongst 23 *de novo* identified proteins, representing significantly more than would be expected by chance (DAPPLE p = 0.0004, p<0.001 correcting for gene size bias). The top-scoring proteins are listed in Table 4.1.

This represents substantial connectivity beyond chance expectation. Furthermore, we obtained a set of *de novo* variants from unaffected siblings of autistic individuals from simplex families and found no significant connectivity amongst 87 genes baring non-synonymous mutations (*direct connectivity* p=.745).

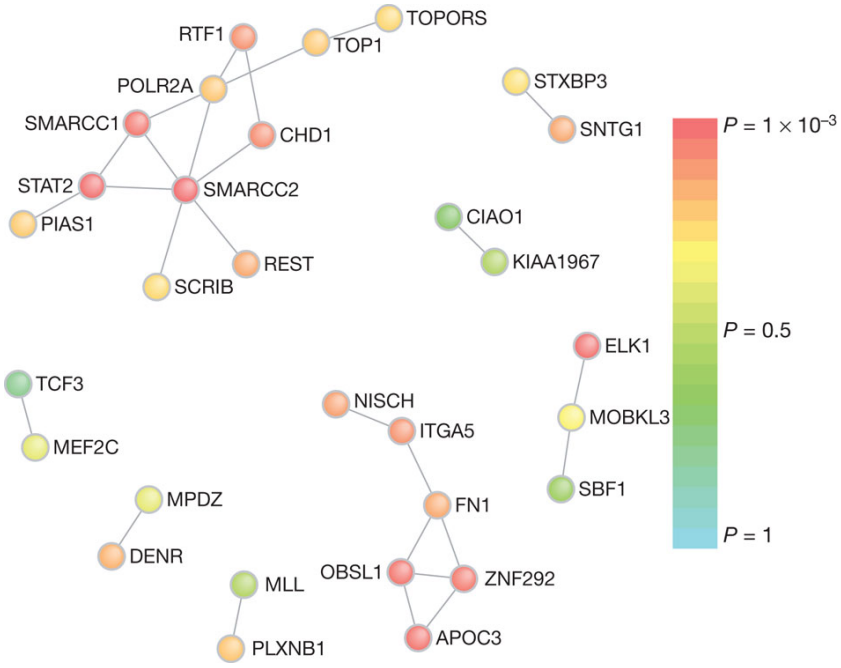


Figure 4.3 Protein–protein interactions for genes with an observed functional *de novo* event. Direct protein connections from InWeb, restricting to genes harboring de novo mutations for DAPPLE analysis. Two extensive networks are identified: the first is centered on SMARCC2 with 12 connections across 11 genes; the second is centered on FN1 with 7 connections across 6 genes. The P-value for each gene having as many connections as those observed is indicated by node color.

We then considered whether the genes highlighted in DAPPLE were known to be haploinsufficient (HI). Since *de novo* variants only affect one copy of the gene, it follows that the functional consequences of a mutation are predicted to be more severe if the gene

is HI. Huang et al. used empirical observations of haploinsufficiency to generate a probabilistic model that assigns HI probabilities to 12,443 genes that they validated using genes known to be implicated in dominant human diseases and mouse knock-out phenotypes[161]. Using a rank-sum test to compare the haploinsufficiency scores of the genes in the *de novo* set that DAPPLE scored highly (DAPPLE p<0.1) to those that were not scored highly, we found an enrichment in HI genes (p=0.0048; the mean HI probability for the DAPPLE proteins was 66.8% whereas the mean overall is 37.6%.)

Table 4.1 DAPPLE and haploinsufficiency scores for *de novo* network proteins. HI scores are probabilities of HI according to Huang et al.

Gene	Chr	DAPPLE p-value	HI score
SMARCC2	Chr12	0.00099975	0.995
STAT2	Chr12	0.00778479	0.466
SMARCC1	Chr3	0.01851351	0.86
APOC3	Chr11	0.01930591	0.255
OBSL1	Chr2	0.01950396	0.159
ELK1	chrX	0.02444871	1
ZNF292	Chr6	0.03056284	0.888
CHD1	Chr5	0.05366016	0.815
RTF1	Chr15	0.06993264	0.595
BRCA2	Chr13	0.076479	0.976
ITGA5	Chr12	0.07667119	0.349

Finally, we tested the direct network in Figure 4.3 for being encoded for by genes co-expressed in neuronal tissue. Using the same expression analysis describe in Chapter 2, we tested each of the 126 tissues in the Benita et al. 2010 dataset for higher-than-expected expression of the network genes [87,108]. Of the 20 tissues that achieved p<0.01, 11 were neuronal (Figure 4.4). This is not expected by chance when compared to tissue enrichment scores of random direct networks built from simulations of *de novo*

variants using the mutation rate model described (p=0.015). Interestingly, most of the other top tissues were related to testes and ovary. These tissues are shown in Table 4.2.

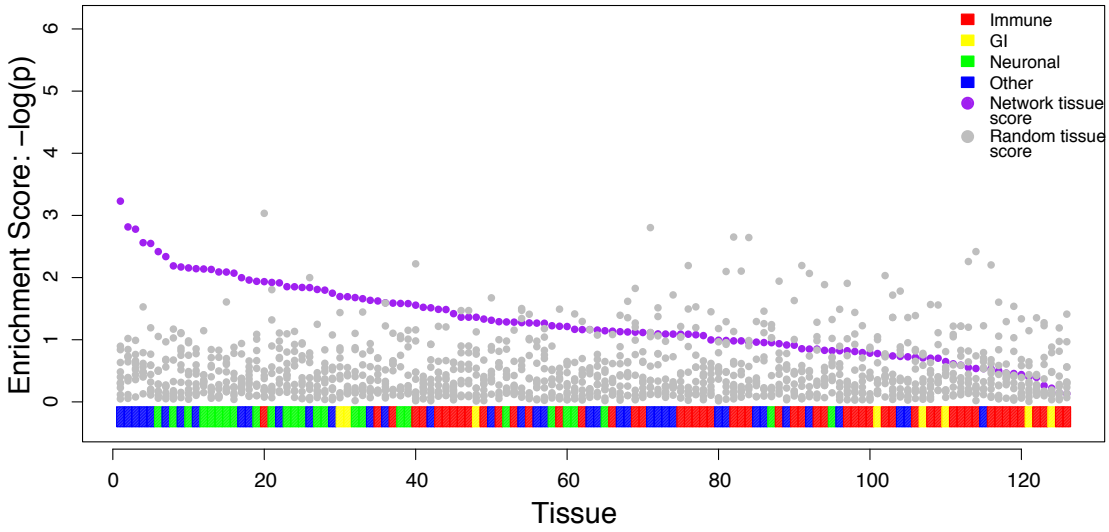


Figure 4.4 Enrichment in neuronally expressed genes in *de novo* network. Using an expression dataset of 126 tissues (described in Chapter 2), we tested the set of 23 network genes for enrichment in tissue expression by comparing their expression values to the rest of the genes in the expression array using a 1-tailed rank-sum test. We found that of the top 22 tissues ($p<0.01$), 11 were neuronal. To test the probability of the top 20 tissues containing ≥ 11 neuronal tissues, we compared the results of the autism network (purple circles) to tissue enrichment scores of random direct networks (grey circles) built from simulations of *de novo* variants using the mutation rate model described (p=0.015).

Table 4.2 Tissue enrichment scores for autism *de novo* network. 11/20 tissues at p<0.01 were neuronal in origin (indicated by a *).

Tissue	Enrichment P-value
Ovary	0.00052348
Testis germ cell	0.001095922
Testis seminiferous tubule	0.00173855
*Dorsal root ganglion	0.002048479
Uterus	0.003299053
Uterus corpus	0.003817946
*Globus pallidus	0.003932643
*Temporal lobe	0.004376913
*Parietal Lobe	0.005247557
Prostate	0.005261165
Heart	0.005543438
*Cingulate cortex	0.005597326
Tongue	0.006592636
*Pituitary	0.006862346
Testis leydig cell	0.007583201
T-regulatory cell	0.008388834
Testis intersitial	0.008567759
*Prefrontal cortex	0.008660865
*Cerebellum	0.008891937
*Superior cervical ganglion	0.00939216

We therefore provide compelling evidence that a subset of the *de novo* variants discovered in autistic patients code for proteins that are non-randomly related to each other through physical interactions. We furthermore provide two additional lines of evidence that this network is biologically relevant, as it is enriched for genes predicted to be HI and the participating genes are co-expressed in neuronal tissues, which may be the relevant cell type in autism.

A subset of the proteins highlighted by the network broadly can be described as involved in chromatin remodeling and transcription. *SMARCC1* and *SMARCC2* achieved nominal DAPPLE p-values of 0.018 and 0.0009, respectively, and were also highly likely

to be HI (86.0% and 99.5%, respectively). These two genes (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, members 1 and 2) are involved in transcriptional activation and repression of genes through chromatin remodeling as part of the SWI/SNF complex. This complex has been shown to be involved in neurogenesis and has been reported to play a role in William's syndrome, a rare autosomal dominant disorder caused by a large chromosomal deletion[162,163]. *STAT2* (p=0.0078) is itself an activator of transcription, though usually in the setting of Type I interferon signaling, and *ELK1* is a transcription factor activated by MAP kinase[164]. Interestingly, chromatin remodeling was also highlighted in a concurrent autism *de novo* study with entirely independent samples and discovered mutations[165].

We therefore provide three lines of evidence that a subset of the genes affected by *de novo* variation may be relevant – they are enriched for proteins that physically bind each other, the connecting proteins are enriched for genes predicted to be haploinsufficient, and the connecting proteins co-expressed in neuronal tissue.

4.3.2 DAPPLE/SEQ: a method to jointly analyze rare variants with PPI data

Section 4.3.1 provides encouraging evidence that protein-protein interaction networks may be of use when studying rare variation, just as has been observed with common-variant associations. We hypothesized that analyzing exome-wide case/control genetic data in the context of large-scale, proteome-wide protein-protein interaction data could reveal previously undiscovered risk genes and pathways. This task is fundamentally distinct from that of *de novo* variation because we typically do not have specific genes to test. Rather, the goal is to look for sub-network enrichment, i.e. pockets of connected proteins that are enriched for association to rare variation.

After consideration of a number of existing sub-network enrichment approaches, we chose to use a data-biased random-walks algorithm described by Komurov et al. (described in section 4.2)[158]. The goal of this approach is to identify pockets of the InWeb database that are enriched for association via rare variation. The method employs a "random walker" who walks throughout the database with each step biased toward more associated nodes. Ultimately, the area of the network around which the walker spent more time should be the most associated and tightly connected cluster. We designed a novel strategy to use this approach to assign PPI-based association scores to genes and called this method Disease Association Protein-Protein Link Evaluator for Sequencing, or DAPPLE/SEQ.

Figure 4.5 shows the workflow of the method. First, all genes are scored for association using the CALPHA burden test for rare variation. Briefly, CALPHA tests for over-dispersion compared to the expected binomial distribution for rare variant counts in cases versus controls (thus being robust to a mix of protective and risk variation)[148]. Genes that are not assigned a score because they bare no variation are automatically assigned a score of p=1. To incorporate protein-protein interaction data, we then calculate a transition matrix as defined in Komurov et al., where the random walker is biased toward transitioning from less associated nodes to more associated nodes. The transition probability between nodes *i* and *j* (p_{ij}) is defined as

$$p_{ij} = \frac{w_j}{\sum_{k \in N_i} w_k}$$

Equation 4.1

where $w = 1/p_{CALPHA}$ and a ceiling is applied, $w_{>.999} = w_{.999}$, to limit the amount of time spent on a highly associated node. As implemented by Komurov et al. and according to the Perron-Frobenius theorem for stochastic matrices, the left eigenvector of the full transition matrix associated with eigenvalue 1 is the vector of stationary probabilities – ie, the final visitation probabilities (which we will refer to herein as ***FVP*** for the full vector of gene-wise final visitation probabilities and FVP_i as the final visitation probability for an individual node)[158]. To further enrich for local pockets of visitation, we look at not only FVP_i for each node but also the sum of its N neighbors:

$$\sum FVP_j = FVP_{N,i}$$

Equation 4.2

These three steps are shown in Figure 4.5A. To correct for the fact that highly connected nodes will be visited more often by chance, we approximated an empirical p-value for each FVP_i and $FVP_{N,i}$ by simulating random genetic data (20,000 simulations of 2,000 cases/2,000 controls each, see section 4.3.3 for description of simulation) and calculating the expected node-wise distribution of FVP_i and FVP_N for all nodes. For each node, the distribution consists of FVP_i and $FVP_{N,i}$ for all nodes of the same binding degree in all null simulations. After comparing FVP_i and $FVP_{N,i}$ to their respective node-wise null distributions, we obtain p_1 and p_2, the empirically derived gene-wise network p-values for FVP_i and $FVP_{N,i}$, respectively (Figure 4.5B). The two empirical p-values are then combined using Brown's method of combining dependent p-values where the test statistic T is assumed to be chi-squared distributed according to the following:

$$T = \frac{-2(\ln(p_1) + \ln(p_2))}{c} \sim \chi^2(df = 2k/c)$$
$$c = \frac{4k + 2\rho(3.25 + .75\rho)}{4k}$$

Equation 4.3

where k=2 and ρ is node-specific and estimated individually for each gene by measuring the correlation between FVP_i and $FVP_{N,i}$ across many simulations.

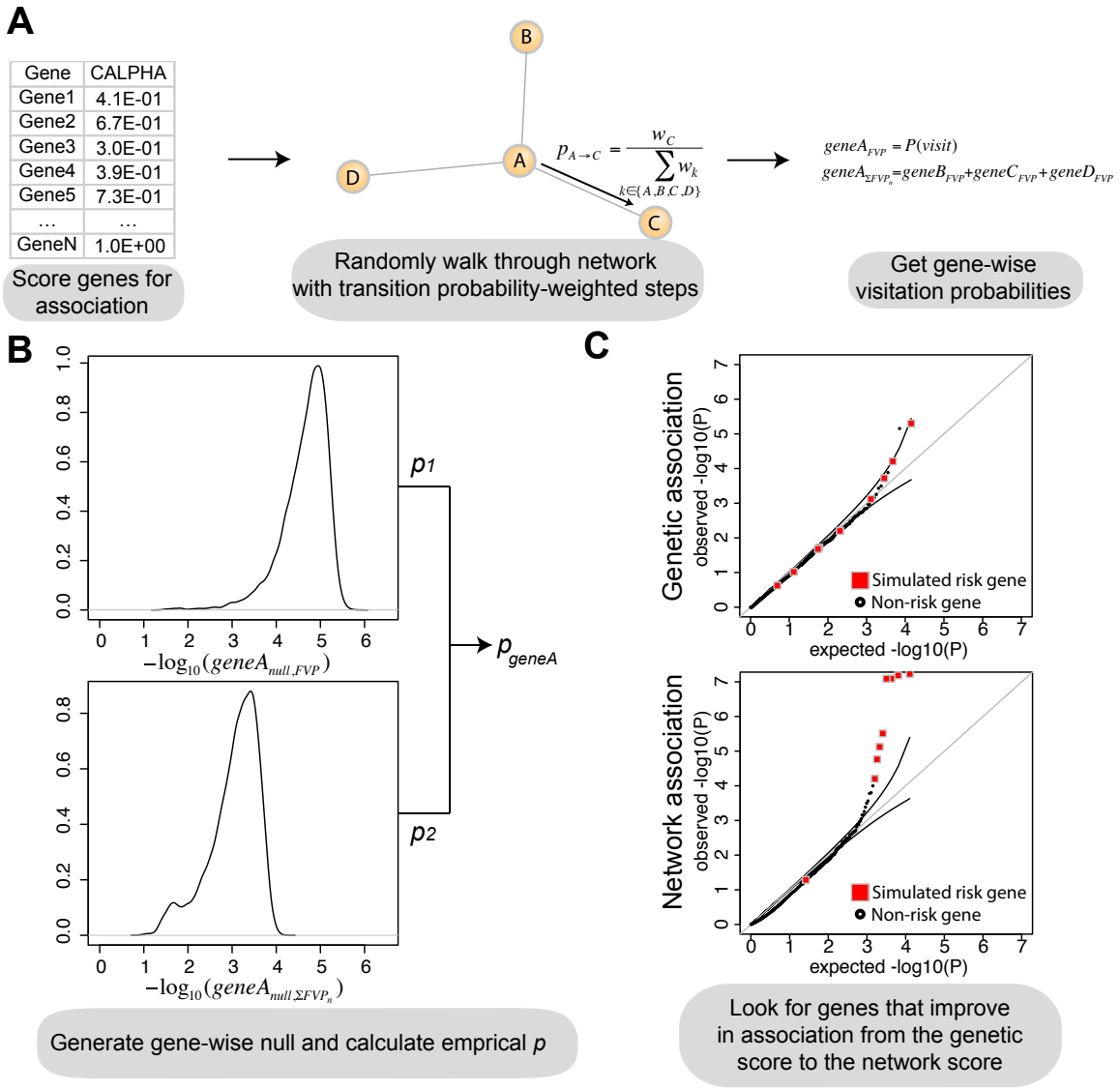


Figure 4.5 Schematic of DAPPLE/SEQ. A. Genes are individually scored for association using the CALPHA method. A transition matrix is calculated by merging these scores with PPI data, and *FVP* and $FVP_{N,i}$ are calculated from that transition matrix. **B.** For each gene *i*, FVP_i and $FVP_{N,I}$ are compared to their expected null distributions based on random genetic data. The resultant empirical p-values are combined to generate a final p_{gene} for each gene. **C.** The distribution of final p-values is then inspected for deviation from null expectation

The final step is to inspect the DAPPLE/SEQ p-value distribution for any genes that are deviating from null expectation (Figure 4.5C). The first Q-Q plot shows the gene-wise p-values using the CALPHA test alone while the second Q-Q plot shows the combined CALPHA and PPI results, where the gene-wise association distribution is now deviating from null expectation and genes containing causal variation (see section 4.3.3 for simulation of risk genes) that are locally interconnected improve in association (red squares). The metric of success for DAPPLE/SEQ is therefore its ability to assign p-values to risk genes that are more significant than the genetic burden test alone.

4.3.3 Simulating risk networks

We tested the method on simulated whole-exome rare variant data where genes in 3 pre-specified networks were assigned causal variation. We used a March 2011 snapshot of sequencing data from an on-going autism case/control sequencing study (52,381 functional variants in 460 cases and 371 controls) to estimate the frequency distribution of rare (<5% MAF) functional alleles (missense, non-sense, splice-site) throughout the exome. The combined allele frequency distribution is shown in Figure 4.6.

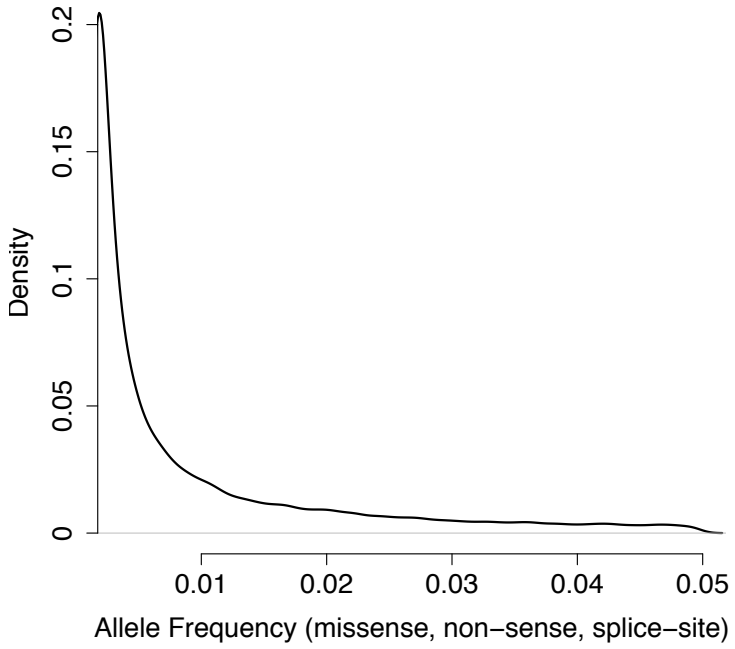


Figure 4.6 Allele frequency distribution for functional alleles. Data from an on-going autism exome sequencing project was used to estimate the allele frequency of 52,381 functional alleles in the human exome (missense, non-sense, splice-site). 1,662 chromosomes in total (460 cases, 371 controls) were assayed on average for each variant. Alleles between MAF 0.0012 (i.e., observed at least twice in the dataset) and 0.05 were included.

To assign risk to variants, we used a liability threshold model with multiplicative risk according to the following biometrical equations. :

$$\mu = a[p^2] + d[2pq] - a[q^2]$$
$$V = (a-\mu)^2 p^2 + (d-\mu)^2 2pq + (-a-\mu)^2 q^2$$
$$a = \frac{Z_{AA} - Z_{aa}}{2}; d = Z_{AA} - a - Z_{Aa}$$
$$Z_G \sim N(1 - prev_G)$$

Equation 4.4

where μ is the population mean of the liability distribution, a is half the distance between the two homozygotes, p is the major allele frequency, q is the minor allele frequency, AA

is homozygous wild-type, aa is homozygous risk, V is the additive genetic variance that the SNP contributes to the phenotypic variance and Z_G represents the liability Z score for a given genotypic category at which the phenotypic cutoff for diagnosis is made. The overall prevalence is assumed to be 5%, and $prev_G$ represents the prevalence within a given genotypic category. We can then apply Bayes theorem to assign genotypes to simulated cases and controls:

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} = \frac{(prev_G)P(G)}{prev}$$

Equation 4.5

where P(G) is the genotype frequency (p^2, 2pq or q^2) and *prev* is the disease prevalence in the population. We simulated rare variants for 2,000 cases and 2,000 controls and used the CALPHA burden test to score each gene for association.

We simulated genetic risk in three biologically relevant networks that are already known to contain disease-causing genes – a network of average binding degree 2.56 consisting of 9 Fanconi Anemia proteins[56,101] that share 23 connections (network 1, Figure 4.7A), a network of average binding degree 1.00 consisting of 10 proteins associated to QT-interval variation[36] that share 10 connections (network 2, Figure 4.7B) and a rheumatoid arthritis[94] network of average binding degree 1.06 consisting of 16 proteins that share 18 connections (network 3, Figure 4.7C). All three networks achieve a significant DAPPLE *direct connectivity* p-value (see Chapter 2). We will herein refer to these networks as networks 1, 2 and 3, respectively.

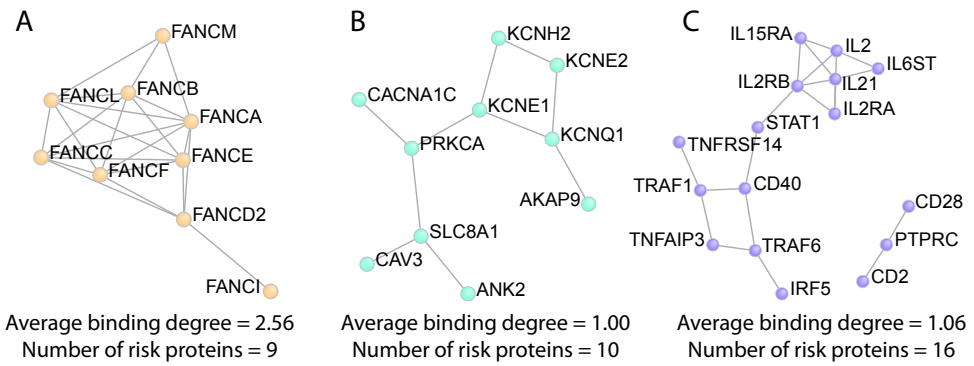


Figure 4.7 Plots and properties of three simulated networks. We chose three networks that represented different properties of sub-networks: network 1 (average binding degree 2.56), network 2 (average binding degree 1.00) and network 3 (average binding degree 1.06).

We varied the amount of additive genetic variance assigned to causal genes between 0.1%, 0.3%, 0.5% and 1%. We saw this range as containing the lower bound for detection (0.1%) and the upper bound for plausibility (1%). 1% additive genetic variance should produce a strong signal of association and is higher than the typical signals observed in GWAS. We therefore assume that low-effect variants discovered through exome-sequencing, even when combined across a gene, will rarely exceed this threshold. For each gene, risk was distributed over 10-20% of the functional (missense, non-sense, splice-site) SNPs therein (or at minimum 1 SNP).

In considering the success of the method, we consider genes that did not achieve $p_{CALPHA} < 2.5e\text{-}6$. These genes will be candidates for follow-up regardless. Since the field is facing a preponderance of results that are above this threshold (hence the need for more innovative methods), we will restrict our results to genes that initially did not achieve such a high score. On average, the proportion of genes that achieve $p_{CALPHA}<2.5e\text{-}6$ is: 0.4% (0.1% additive genetic variance), 5.6% (0.3% additive genetic variance), 18.6% (0.5% additive genetic variance), 39.6% (1% additive genetic variance).

4.3.4 Running DAPPLE/SEQ on simulated risk networks

The results of running DAPPLE/SEQ on networks 1, 2 and 3 are summarized in Figure 4.8 and Figure 4.9. First, we asked whether the method was powered to detect significance at $p_{DAPPLE/SEQ}$<2.5e-6. We found that in all three networks, DAPPLE/SEQ assigned genome-wide significant scores to risk proteins previously at p_{CALHA}>2.5e-6. The percent of such proteins improved consistently as the average additive genetic variance assigned to risk genes was increased from 0.1% to 1.0%.

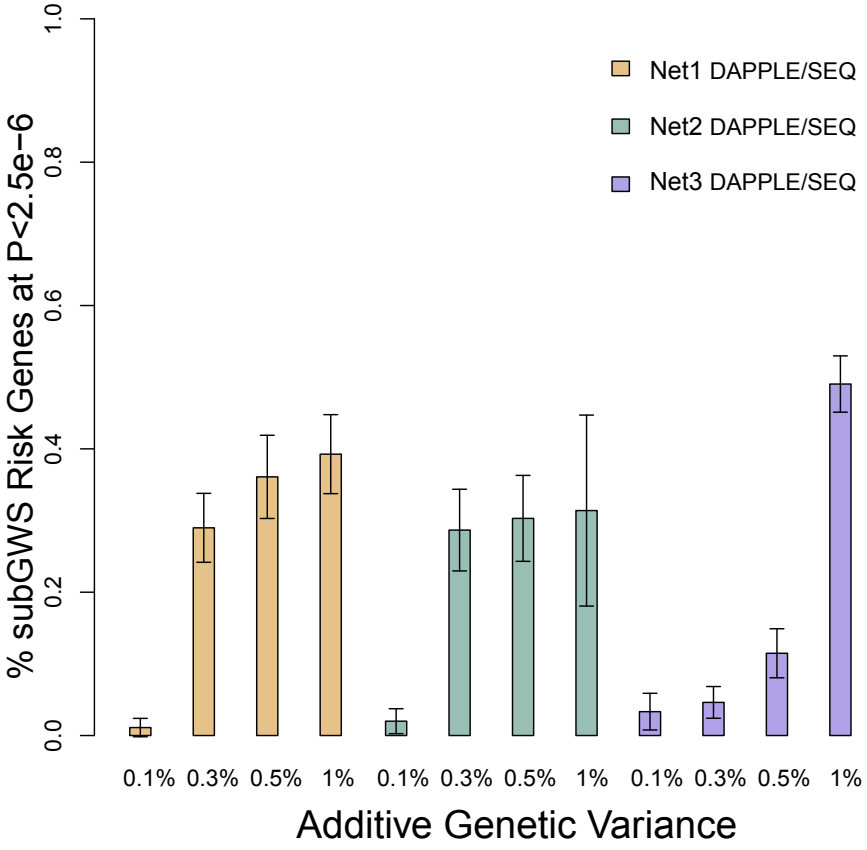


Figure 4.8 Percent of risk genes that rise to genome-wide significance using DAPPLE/SEQ. After running DAPPLE/SEQ on ~50 simulated genetic datasets for all three risk networks (2000 cases, 2000 controls for each simulation), we counted the number of risk genes that earned a score that passes the Bonferroni corrected cutoff of 2.5e-6. 95% confidence intervals are shown.

We then relaxed our threshold and asked whether more proteins were assigned scores in the tail of the DAPPLE/SEQ association distribution as compared to CALPHA alone. Figure 4.9 shows the percentage of risk genes (originally at p_{CALPHA}>2.5e-6) that achieved p_{CALPHA}<1e-4 as compared to the percentage of risk genes that achieved $p_{DAPPLE/SEQ}$<1e-4 (i.e., the true positive rate at p<1e-4). When compared to the same cutoff using CALPHA alone, the true-positive rate for DAPPLE/SEQ at p<1e-4 is improved significantly in most cases if the additive genetic variance is above 0.1% (Wilcox rank-sum two-tailed p-values for CALPHA vs. DAPPLE/SEQ at 0.1%, 0.3%, 0.5% and 1.0% additive genetic variance, respectively, are: *network 1*: 0.0715, 9.32e-13, 2.01e-14, 2.19e-16; *network 2*: 0.0723, 4.53e-8, 1.89e-6, 0.0036; *network 3*: 0.257, 0.515, 0.00373, 2.87e-13).

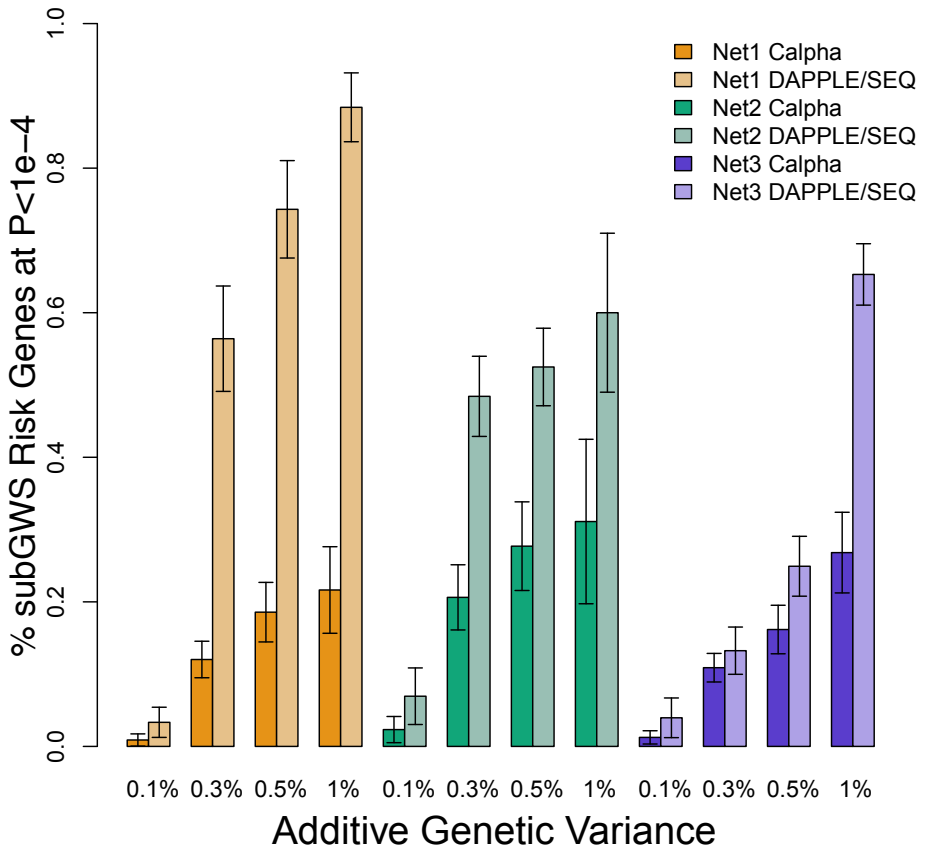


Figure 4.9 DAPPLE/SEQ improves true positive rate at p<1e-4. We compared the true positive rate at p<1e-4 for the CALPHA test versus DAPPLE/SEQ. We found that for additive genetic variances greater than 0.1%, the method significantly improves the ability to detect risk genes (for 0.1%, 0.3%, 0.5% and 1% additive genetic variance, rank-sum comparison p-values are *network 1*: 0.0715, 9.32e-13, 2.01e-14, 2.19e-16; *network 2*: 0.0723, 4.53e-8, 1.89e-6, 0.0036; *network 3*: 0.257, 0.515, 0.00373, 2.87e-13). 95% confidence intervals are shown.

It is unlikely that the underlying disease etiology for complex traits consists of a single, small network. Afterall, what we have learned from GWAS is that complex traits are extremely polygenic. We were therefore interested in the behavior of DAPPLE/SEQ if the disease etiology were due two separate mechanisms. We simulated risk simultaneously in network 1 and network 2 (19 proteins in total). We were encouraged to

find the same general trends. First, risk genes previously at p_{CALPHA} > 2.5e-6 became genome-wide significant, though the percent of risk genes at this cutoff leveled off after 0.3% additive genetic variance (Figure 4.10, see discussion for why this might be).

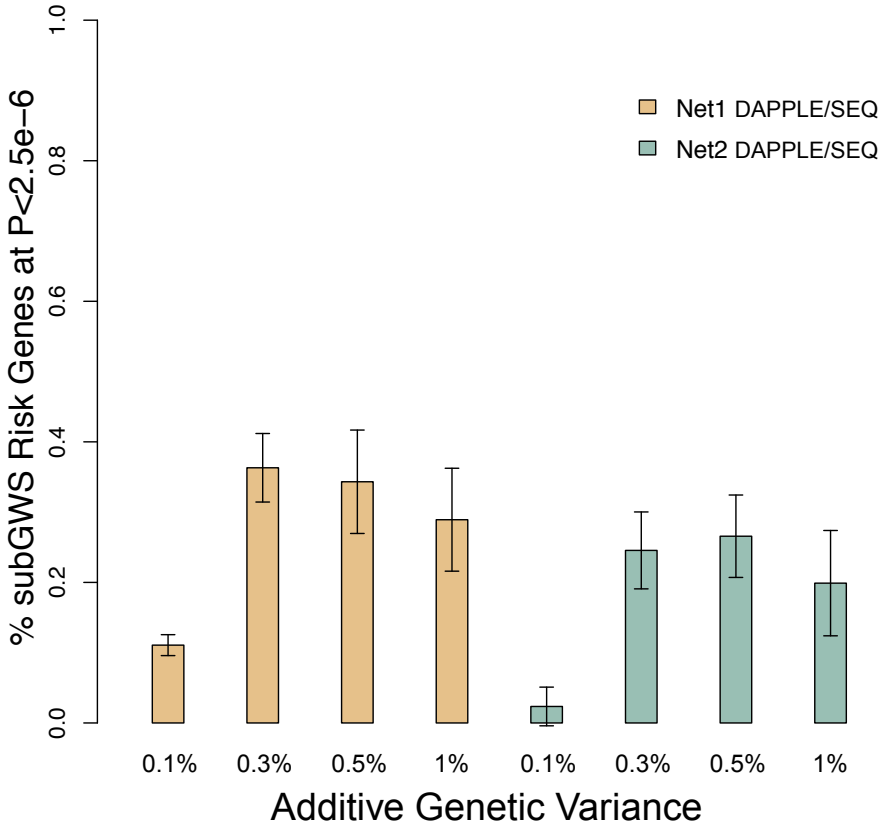


Figure 4.10 Percent of risk genes that rise to genome-wide significance after introduction of PPI data for joint networks 1 and 2. We simulated risk on networks 1 and 2 simultaneously for ~50 datasets (2000 cases, 2000 controls for each simulation) and ran DAPPLE/SEQ. We then counted the number of risk genes that earned a score that passes the Bonferroni corrected cutoff of 2.5e-6. 95% confidence intervals are shown.

Second, a consistent increase in genes achieving $p_{DAPPLE/SEQ}$<1e-4 was also observed for the joint network (Figure 4.11). When compared to the same cutoff using CALPHA alone, the true-positive rate for DAPPLE/SEQ at p<1e-4 is improved

significantly in most cases (Wilcox rank-sum two-tailed p-values for CALPHA vs. DAPPLE/SEQ at 0.1%, 0.3%, 0.5% and 1.0% additive genetic variance, respectively, are: *network 1*: 1.31e-10, 3.44e-10, 3.5e-9, 6.78e-9; *network 2*: 0.435, 3.95e-5, 0.00201, 0.00494).

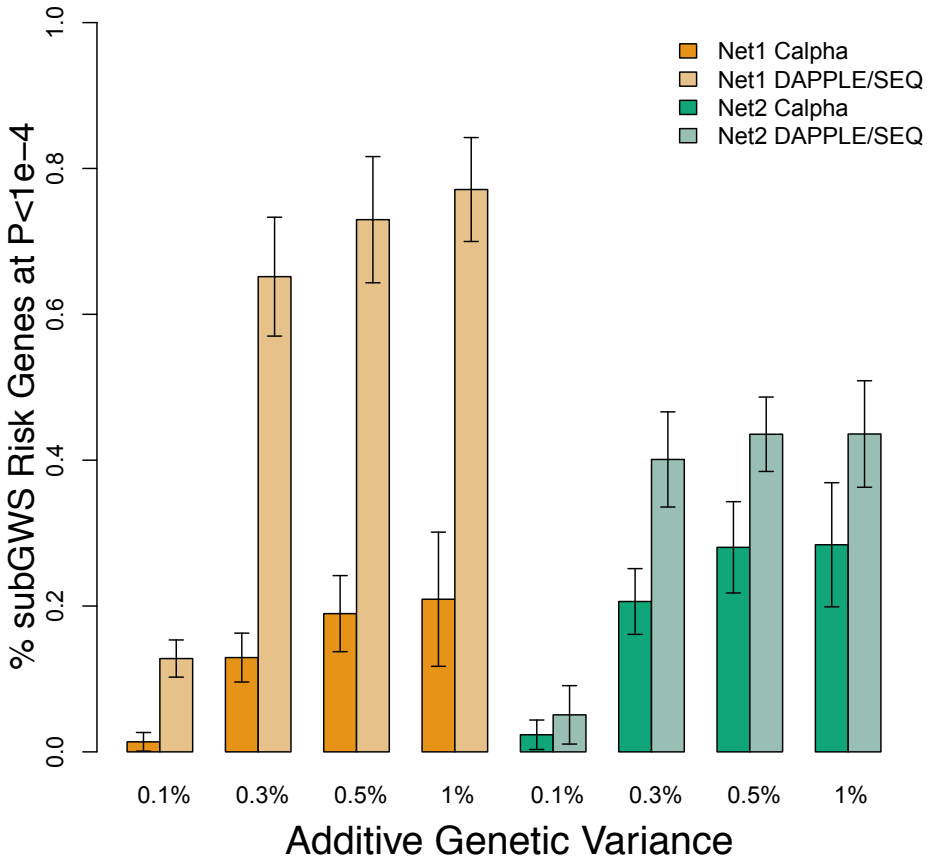


Figure 4.11 DAPPLE/SEQ improves p-values for risk genes in a joint model with networks 1 and 2 are simultaneously associated. We simulated risk simultaneously in networks 1 and 2 and compared the true positive rate at p<1e-4 for the CALPHA test versus DAPPLE/SEQ. When compared to the same cutoff using CALPHA alone, the true-positive rate for DAPPLE/SEQ at p<1e-4 is improved significantly in most cases (Wilcox rank-sum two-tailed p-values for CALPHA vs. DAPPLE/SEQ at 0.1%, 0.3%, 0.5% and 1.0% additive genetic variance, respectively, are: *network 1*: 1.31e-10, 3.44e-10, 3.5e-9, 6.78e-9; *network 2*: 0.435, 3.95e-5, 0.00201, 0.00494). 95% confidence intervals are shown.

The false positive rates for all 4 risk networks (network 1, network 2, network 3 and joint networks 1&2) are shown in Figure 4.12. The reason behind a non-zero false positive rate is that sometimes, associated sub-networks will cause close PPI neighbors to become highly significant. It is therefore important to consider CALPHA results jointly with DAPPLE/SEQ results in evaluating candidates for follow-up.

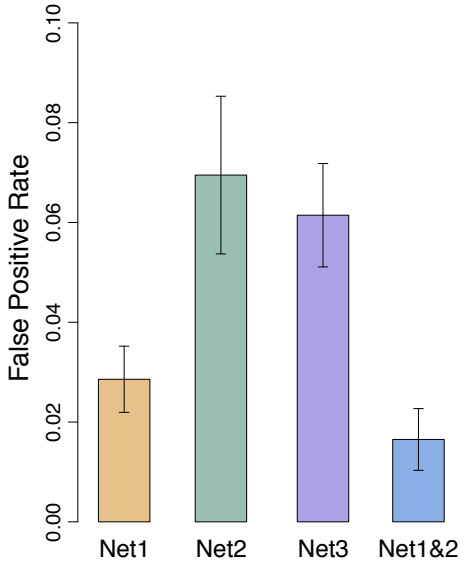


Figure 4.12 False positive rates at p<2.5e-6 for DAPPLE/SEQ on 4 risk networks. The percent of non-risk genes achieving $p_{DAPPLE/SEQ}$<2.5e-6 for all 4 networks is shown. 1 standard deviation is plotted.

We then asked whether the method controls for binding degree bias in the data. We found that it may be conservative with respect to nodes of high binding degree. As discussed in chapter 2, the InWeb database (and PPI databases in general) is affected by publication bias: well studied proteins will appear to be more connected than others, even if they are not so in the true set of biological connections. As shown in Figure 4.13, there is a significant negative correlation (r=-0.06, p=8.6e-10) and genes of high binding degree are rarely significant, suggesting that DAPPLE/SEQ is conservative; that is,

highly connected nodes are less likely to achieve significance. This likely explains the more modest effect observed with network 3, whose mean global binding degree (the average binding degree in the InWeb database) is 73.6, while for networks 1 and 2 it is 25.6 and 45.5.

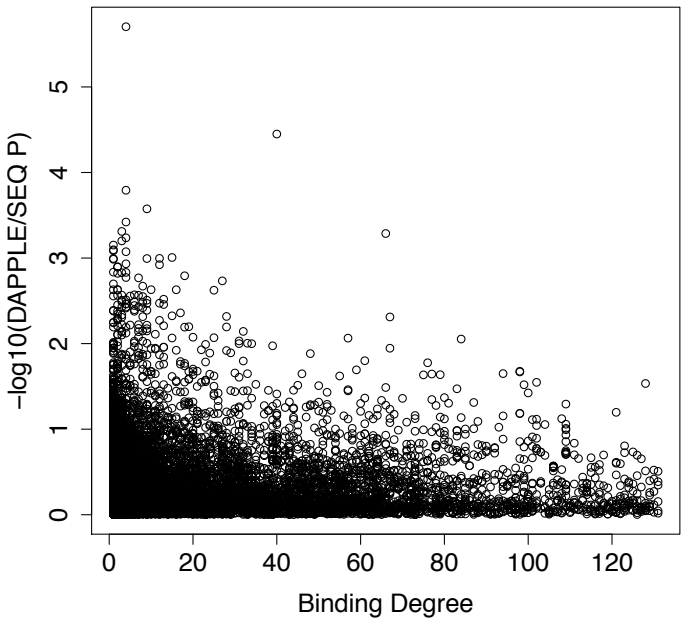


Figure 4.13 Correlation between DAPPLE/SEQ p-value and protein binding degree. A representative simulation (0.3% additive genetic variation, network 1) is plotted (proteins at > 3 standard deviations away in binding degree were removed). r=-0.06, p=8.6e-10.

4.4 Discussion

In this chapter, we have provided preliminary evidence that considering functional relationships between genes when analyzing rare variants (both *de novo* and inherited) can highlight risk genes that would have been otherwise missed. The results presented here are extremely promising to the field, as investigators are now looking for rare variation that influences risk to disease but are uniformly encountering significant analytic challenges[150]. Based on early results, most have found that the current

methods for association analysis (both SNP-wise and gene-wise via burden tests) are underpowered due to the extreme rarity of these types of genomic events. For diseases where risk is distributed over functional networks – which is likely the case for many heritable complex traits, as we have shown in previous chapters – introducing protein-protein interaction data to the analysis might offer a novel route to identifying risk genes and processes.

The network assembled from the autism *de novo* variants is extremely promising. In sporadic cases of autism (simplex families), one hypothesized genetic etiology is that *de novo* point mutations make contribute to risk in some cases, similar to sporadic CNVs [55]. While a large burden of *de novo* variation was not necessarily observe beyond a small excess of non-sense mutations, we show that the genes baring *de novo* variation in autism patients is significantly enriched for proteins that physically connect to each other. We show that the network genes are co-expressed in brain and that the genes implicated suggest chromatin and transcriptional regulation, as has been recently found by another group[165]. This type of analysis therefore might be promising in terms of identifying particular genes on which to follow up. Moreover, this network can now serve as a candidate for future case/control focus, as mechanisms affected by *de novo* variants that contribute to disease may be similar to those affected by inherited variants.

We have also described a method to extend this concept to rare variants discovered through case/control sequencing. Similar to the early days of GWAS, exome sequencing studies so far have found very little evidence for strong associations. Almost certainly, increasing power should yield more results in the future. However, it may be the case for rare variants that distinguishing them from background variation is so difficult that joint

consideration of biological process in genetic association studies will be required. While the results are based only o n simulated data, they serve as a principled test of the method (DAPPLE/SEQ) and proof that detecting local areas of PPI connectivity associated to rare variation is possible, even when the additive variance explained by a gene is considerably low.

Interestingly, DAPPLE/SEQ performs much better on networks 1 and 2 than it does for network 3. We hypothesize that the method is limited in its detection of association at nodes of unusually high binding degree in the InWeb database, likely due to the step that controls for visitation bias due to degree. Indeed, Figure 4.13 shows that nodes of higher degree have systematically lower association scores. In future applications of this approach, solutions to this problem should be considered, such as down-sampling of edges for highly-connected nodes. It is also interesting to note that DAPPLE/SEQ is slightly power-limited in scenarios of multiple risk mechanisms as additive genetic variance per gene increases (Figure 4.10). The likely driving force is that the method is competitive: nodes compete with each other for visitation. In scenarios of multiple disconnected mechanisms, each one draws visitation time from the next, which might impair signal to noise detection. Nonetheless, the approach still out-competes tests of genetic association alone.

It is important to underscore that these results are preliminary. The inherent limitation to simulations is that it is computationally unfeasible to test the full range of possible parameters. Future work will include varying gene-wise assigned risk, trait heritability, network size, network connectivity and percent of network affected. In addition, other functional connectivity datasets can be used – for example, co-expression

datasets are well suited for this application. Nonetheless, the results described here should serve as promising evidence that the prospective consideration of the functional relationships between genes can significantly improve the discovery of rare variants influencing disease.

4.5 Acknowledgements

Contributed de novo sequencing data:	Benjamin Neale, Kaitlin Samocha, Mark J Daly
Performed tailored DAPPLE de novo analysis:	Elizabeth J Rossin
Generated *de novo* simultions:	Shamil Sunyaev, Jared Maguire
Performed expression and haploinsufficiency analysis:	Elizabeth J Rossin
Overall idea, concept and project coordination for DAPPLE/SEQ:	Elizabeth J Rossin, Menachem Fromer, Chris Cotsapas, Mark J Daly
Conceived and designed DAPPLE/SEQ method:	Elizabeth J Rossin, Menachem Fromer, Mark J Daly
Carried out DAPPLE/SEQ analyses and simulations:	Elizabeth J Rossin

5 Discussion

This thesis describes three approaches to combine genetic and proteomic data. In Chapter 1, we describe an *in silico* PPI analysis tool to test loci associated to complex traits for harboring genes that are significantly related through physical interactions. Abundant evidence is provided to show that this method, called "DAPPLE", is able to detect enriched connectivity in loci associated to two autoimmune diseases, prioritize genes in large loci and propose new genes for follow-up based on their connectivity to the disease network. We then explore the use of directed experiments in discovering PPIs by focusing on the genetics of QT-interval variation. We show that the heart-specific protein complexes of 5 known long-QT syndrome proteins are not only highly enriched for proteins near common variants associated to QT-interval variation but also able to point to novel genes previously not known to affect the QT-interval. We extend this concept to rare variation in Chapter 4, which due to its infrequent nature requires novel statistical methodological development for analysis. Using a case example of *de novo* variation in autism, we show that DAPPLE can successfully highlight *de novo* variants that may be relevant to autism even if no standard statistical methodology can be employed to identify specific genes on which to follow up. Finally, we implement and test a tool ("DAPPLE/SEQ") to analyze rare inherited genetic variation in the context of protein-protein interaction data. Using simulated risk variation, we show that DAPPLE/SEQ significantly improves the ability to detect rare risk variants that are distributed over a network of interacting proteins.

Since being published, DAPPLE has been widely used by the community and therefore empirically tested as to its applicability to a broad spectrum of complex traits. In the first year, it has been used 1,500 times by 350 unique users across the world. For example, Raj and colleagues found that Alzheimer's disease susceptibility loci under natural selection form significant interaction networks[166]; Cotsapas and colleagues showed discrete networks associated with clusters of shared autoimmune susceptibility loci[167]; Morris and colleagues identified a network underlying new loci associated to Type 2 Diabetes [manuscript accepted at *Nature Genetics*]; and Irvin and colleagues identified a significant network underlying genes associated to subcutaneous adipose tissue distribution in HIV-infected men[168]. In our own work, we have observed significant underlying connectivity in susceptibility loci for Crohn's disease, rheumatoid arthritis, blood-lipid levels, height, QT-interval variation, autism (via *de novo* point mutations), schizophrenia[169] and multiple sclerosis, among others. The results therefore seem to extend beyond the particular phenotypes studied here, and more generally provide evidence for the concept that loci associated to complex traits code for proteins that physically interact with one another.

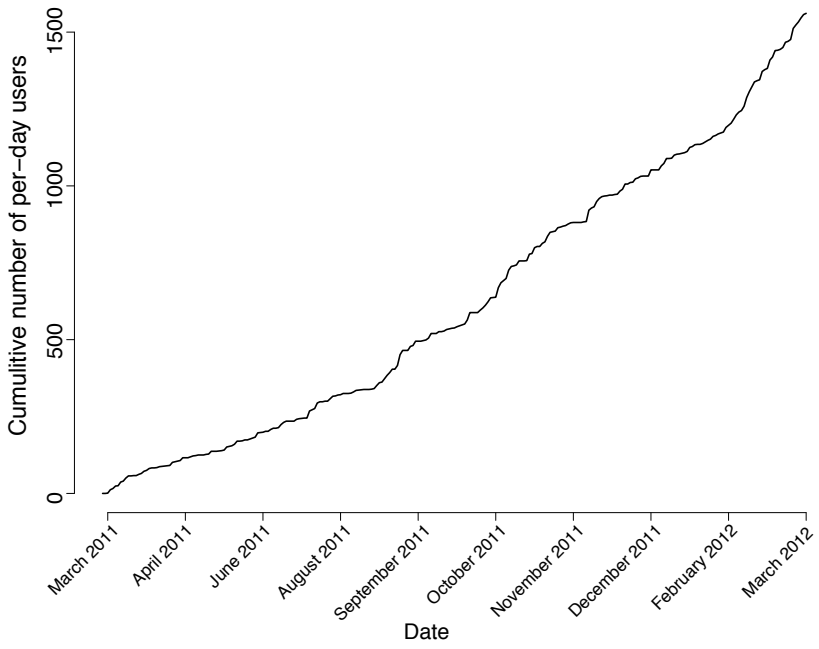


Figure 5.1 Cumulative per-day users of DAPPLE

The concept of applying protein-protein interaction networks to genes of interest is not new, but only recently has it begun to be rigorously employed to human genetics. Many groups have also recently integrated proteomics and genetics in creative ways to highlight mechanisms underlying complex traits. Wang et al. recently created a 3-dimensional structural interface map of human protein-protein interactions, and showed that disease-causing mutations are more likely to be at the interaction interface of the associated protein, with location within the interface giving specificity to the particular disease[170]. Dutkowski and Ideker showed that better predictive models can be made in breast and brain cancer metastasis if one combines genetic mutations with the protein networks that are affected[171]. Sang et al. used purification of ciliopathy protein complexes to link different ciliopathies to specific mechanisms as well as to discover new genes causal of ciliopathy in humans[172]. Though the idea of deciphering the protein-

binding partners of genes of interest is an old approach, the formal integration of large-scale networks with human complex trait genetics has been receiving more and more attention over the last decade.

The future of integrative *in silico* PPI and genetic network analysis will likely include more specific and multifaceted snapshots of cellular processes. For example, in Chapter 2 and 4 we overlay expression data onto CD, RA and autism networks to make sure that the interactions identified were between proteins expressed in the same tissue and ensure that the highlighted tissue makes sense in the context of disease (though this does not necessarily have to be the case and could offer insight into unexpected biology). However, it is likely that a more formal integration of expression data with PPI data will be powerful and will allow testing of tissue-specific networks, rather than *post hoc* analysis. In addition to co-expression, other types of connectivity information – such as co-regulation and genetic interactions – is going to be useful and represents some of the main foci of systems biology today [173–175]. Furthermore, interactions do not necessarily need to be represented as binary: here, we binarize the InWeb database for ease of analysis; however, using the raw confidence scores might improve analysis capabilities by calculating weighted distances between nodes rather than restricting to one- and two- degree relationships.

In addition, the future will likely also include dynamic snapshots of cellular processes that are relevant to the phenotype being studied. While some interactions are stable across different cellular stresses, it is likely that the topology of the underlying interaction landscape changes as perturbations (genetic, pharmacologic, etc) arise. Therefore, it will become important to construct “differential” networks, where edges are

qualitatively characterized by how they change under specified conditions. This type of analysis would be alike to expression analysis, where measuring differential expression with condition is considered the norm[176].

What has become clear from extensive exploration of the public PPI data, however, is that we cannot rely solely on public databases: directed *in vivo* experiments are warranted. In chapter 3, we focused on 5 proteins and elucidated their heart-specific interaction partners. While the members of the complexes significantly overlapped with interactions documented in the literature, there were many interactions that are not only novel but involve proteins that are relevant to disease and affect the QT-interval when tested in Xenapus oocytes and zebrafish. Looking forward, it will likely be extremely powerful to carry out gene-specific and tissue-specific PPI experiments rather than relying only on public databases. For example, Sakai et al. recently conducted a yeast-2-hybrid screen of 26 autism-associated genes and found numerous novel connections, including the ability to predict new genes that when tested in independent samples were affected by CNVs in autistic individuals[155]. The two approaches (public and directed) do not have to be mutually exclusive but rather can be cooperative, whereby public databases help generate testable hypotheses and are in turn dynamically augmented by specific experiments.

Furthermore, as proteomic technologies continue to advance, proteome-wide complex purifications will hopefully become feasible. This will allow for the unbiased assaying of the full protein-protein interactome without relying on the literature, which suffers from publication bias. Though no approach will be truly unbiased (since certain proteins will naturally be easier to assay than others, certain tissues will be easier to

culture, etc), it will fill in missing pieces of the interactome that are currently less well-studied. These pieces likely will provide novel insights into regions of the genome that are definitively associated to disease but poorly annotated.

The ultimate promise of network-based interpretation of disease is that it might help steer therapeutic design. For example, Berger et al. showed that the sub-network surrounding long-QT proteins is enriched for FDA-approved drug targets that *cause* prolongation of the QT-interval, a known side effect of many drugs[134]. They then went on to predict and validate drugs that affect the QT-interval, illustrating the general principle that biological networks built from disease-associated genes can point to new proteins relevant to therapy. While this study focused on adverse events, a logical next step is using networks to point to proteins as novel targets for disease or that are better targets than the current ones. The current target-based drug discovery platforms are poor at predicting drug efficacy and the full spectrum of side-effects, which is one of the reasons that most drugs do not survive the drug development process[177,178]. Moving away from target-based design toward network-centered design may help increase the efficiency of the therapeutic design process because a more complete consideration of the biological context of a drug target will take place.

Hopefully, we have provided convincing evidence in this thesis that the field of human genetics is ready to start projecting its findings onto networks of gene-gene relationships. Studying variants associated to disease in the context of the biological systems in which they exert their effect is of utmost importance to translating genetic findings to function and ultimately to medical relevance. The entire goal of human genetics is, afterall, to unveil the pathophysiologic mechanisms that are currently

unknown. To achieve this goal, it will be very powerful to include in our analyses the fact that genes and their protein products do not act alone; they are part of complex networks of proteins that together produce the phenotypes that we study in human genetics.

Bibliography

1. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. 33 Suppl: 228–237. Accessed 21 April 2012.

2. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881–888. Accessed 11 October 2010.

3. Sturtevant A (1913) The Linear Arrangement of Six Sex-linked Factors in Drosophila, as Shown by Their Mode of Association. Journ Exp Zoo 14: 43–59.

4. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32: 314–331. Accessed 14 March 2012.

5. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234–238. Accessed 14 March 2012.

6. Online Mendelian Inheritance in Man. Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/omim. Accessed 14 March 2012.

7. Klein J, Sato A (2000) The HLA system. First of two parts. N. Engl. J. Med. 343: 702–709. Accessed 14 March 2012.

8. Strittmatter WJ, Roses AD (1996) Apolipoprotein E and Alzheimer's disease. Annu. Rev. Neurosci. 19: 53–77. Accessed 14 March 2012.

9. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The Structure of Haplotype Blocks in the Human Genome. Science 296: 2225–2229. Accessed 15 March 2012.

10. Consortium TIH (2005) A haplotype map of the human genome. Nature 437: 1299. Accessed 15 March 2012.

11. Ragoussis J (2009) Genotyping Technologies for Genetic Research. Annual Review of Genomics and Human Genetics 10: 117–133. Accessed 15 March 2012.

12. de Bakker PIW, Neale BM, Daly MJ (2010) Meta-analysis of genome-wide association studies. Cold Spring Harb Protoc 2010: pdb.top81. Accessed 21 September 2010.

13. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58. Accessed 21 September 2010.

14. McGovern DPB, Gardet A, Törkvist L, Goyette P, Essers J, et al. (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat. Genet 42: 332–337. Accessed 24 October 2010.

15. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/19430480. Accessed 19 March 2010.

16. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet 40: 955–962. Accessed 27 February 2010.

17. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, et al. (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat. Genet 41: 1330–1334. Accessed 19 March 2010.

18. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118–1125. Accessed 3 December 2010.

19. McCarthy MI, Zeggini E (2009) Genome-wide association studies in type 2 diabetes. Curr Diab Rep 9: 164–171. Accessed 18 August 2010.

20. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, et al. (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet 42: 579–589. Accessed 18 August 2010.

21. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet 40: 638–645. Accessed 18 August 2010.

22. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet 40: 161–169. Accessed 3 August 2010.

23. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, et al. (2008) Many sequence variants affecting diversity of adult human height. Nat. Genet 40: 609–615. Accessed 3 August 2010.

24. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40: 575–583. Accessed 28 July 2010.

25. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet 42: 565–569. Accessed 23 September 2010.

26. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science 316: 889–894. Accessed 14 March 2012.

27. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 106: 9362–9367. Accessed 9 February 2012.

28. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466: 714–719. Accessed 24 September 2010.

29. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat. Genet 40: 1107–1112. Accessed 23 April 2010.

30. Kuballa P, Huett A, Rioux JD, Daly MJ, Xavier RJ (2008) Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. PLoS ONE 3: e3391. Accessed 27 February 2010.

31. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707–713. Accessed 16 March 2012.

32. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308: 385–389. Accessed 15 March 2012.

33. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752. Accessed 12 February 2012.

34. Ruderfer DM, Kirov G, Chambert K, Moran JL, Owen MJ, et al. (2011) A family-based study of common polygenic variation and risk of schizophrenia. Mol Psychiatry 16: 887–888. Accessed 12 February 2012.

35. Marks D, Thorogood M, Neil HAW, Humphries SE (2003) A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. Atherosclerosis 168: 1–14. Accessed 16 March 2012.

36. Newton-Cheh C, Eijgelsheim M, Rice KM, de Bakker PIW, Yin X, et al. (2009) Common variants at ten loci influence QT interval duration in the QTGEN Study. Nat. Genet 41: 399–406. Accessed 2 July 2011.

37. Pfeufer A, Sanna S, Arking DE, Müller M, Gateva V, et al. (2009) Common variants at ten loci modulate the QT interval duration in the QTSCD Study. Nat. Genet 41: 407–414. Accessed 2 July 2011.

38. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838. Accessed 15 June 2012.

39. Hirschhorn JN, Gajdos ZKZ (2011) Genome-wide association studies: results from the first few years and potential implications for clinical medicine. Annu. Rev. Med. 62: 11–24. Accessed 15 March 2012.

40. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224. Accessed 22 April 2010.

41. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet 4: e1000214. Accessed 5 March 2010.

42. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, et al. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet. 85: 13–24. Accessed 16 March 2012.

43. Moskvina V, O'Dushlaine C, Purcell S, Craddock N, Holmans P, et al. (2011) Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. Genet. Epidemiol. 35: 861–866. Accessed 16 March 2012.

44. Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet 6. Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20714348. Accessed 11 October 2010.

45. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A 102: 15545–15550. Accessed 16 March 2010.

46. Pedroso I, Lourdusamy A, Rietschel M, Nöthen MM, Cichon S, et al. (2012) Common Genetic Variants and Gene-Expression Changes Associated with Bipolar Disorder Are Over-Represented in Brain Signaling Pathway Genes. Biological Psychiatry Available at: http://www.ncbi.nlm.nih.gov/pubmed/22502986. Accessed 29 April 2012.

47. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30. Accessed 6 December 2010.

48. D'Eustachio P (2011) Reactome knowledgebase of human biological pathways and processes. Methods Mol. Biol 694: 49–61. Accessed 6 December 2010.

49. Bartel DP (2009) MicroRNAs: Target Recognition and Regulatory Functions. Cell 136: 215–233. Accessed 21 April 2010.

50. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854. Accessed 22 November 2010.

51. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, et al. (2011) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol. Psychiatry 16: 286–292. Accessed 16 March 2012.

52. Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. Cell 147: 57–69. Accessed 8 February 2012.

53. Conrad DF, Keebler JEM, DePristo MA, Lindsay SJ, Zhang Y, et al. (2011) Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43: 712–714. Accessed 9 February 2012.

54. Aschard H, Qiu W, Pasaniuc B, Zaitlen N, Cho MH, et al. (2011) Combining effects from rare and common genetic variants in an exome-wide association study of sequence data. BMC Proceedings 5 Suppl 9: S44. Accessed 12 April 2012.

55. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372. Accessed 23 October 2010.

56. Moldovan G-L, D'Andrea AD (2009) How the fanconi anemia pathway guards the genome. Annu. Rev. Genet 43: 223–249. Accessed 13 May 2010.

57. Beales PL, Badano JL, Ross AJ, Ansley SJ, Hoskins BE, et al. (2003) Genetic interaction of BBS1 mutations with alleles at other BBS loci can result in non-Mendelian Bardet-Biedl syndrome. Am. J. Hum. Genet 72: 1187–1199. Accessed 12 March 2010.

58. Lim J, Hao T, Shaw C, Patel AJ, Szabó G, et al. (2006) A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration. Cell 125: 801–814. Accessed 12 March 2010.

59. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am. J. Hum. Genet 78: 1011–1025. Accessed 12 March 2010.

60. Bader GD, Betel D, Hogue CWV (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res 31: 248–250. Accessed 25 October 2010.

61. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572–574. Accessed 25 October 2010.

62. Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol 25: 309–316. Accessed 27 February 2010.

63. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Res 38: D525–531. Accessed 6 December 2010.

64. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database--2009 update. Nucleic Acids Res 37: D767–772. Accessed 25 October 2010.

65. Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, et al. (2008) MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res 36: D196–201. Accessed 25 October 2010.

66. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30: 303–305. Accessed 25 October 2010.

67. K|[ouml]|cher T, Superti-Furga G (2007) Mass spectrometry|[ndash]|based functional proteomics: from molecular machines to protein networks. Nature Methods 4: 807. Accessed 12 April 2012.

68. Fields S, Song O (1989) A novel genetic system to detect protein–protein interactions. 340: 245. Accessed 1 March 2012.

69. Lievens S, Lemmens I, Tavernier J (2009) Mammalian two-hybrids come of age. Trends in Biochemical Sciences 34: 579–588. Accessed 1 March 2012.

70. Sardiu ME, Washburn MP (2011) Building protein-protein interaction networks with proteomics and informatics tools. J. Biol. Chem. 286: 23645–23651. Accessed 1 March 2012.

71. Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, et al. (2008) An experimentally derived confidence score for binary protein-protein interactions. Nature Methods 6: 91. Accessed 1 March 2012.

72. Yu H, Braun P, Yıldırım MA, Lemmens I, Venkatesan K, et al. (2008) High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science 322: 104–110. Accessed 2 March 2012.

73. Kaake RM, Wang X, Huang L (2010) Profiling of protein interaction networks of protein complexes using affinity purification and quantitative mass spectrometry. Mol. Cell Proteomics 9: 1650–1665. Accessed 4 March 2012.

74. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, et al. (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nature Biotechnology 17: 1030. Accessed 4 March 2012.

75. Chang I (2006) Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. PROTEOMICS 6: 6158–6166. Accessed 4 March 2012.

76. Vermeulen M, Hubner NC, Mann M (2008) High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. Current Opinion in Biotechnology 19: 331–337. Accessed 12 April 2012.

77. Shoemaker BA, Panchenko AR (2007) Deciphering Protein–Protein Interactions. Part I. Experimental Techniques and Databases. PLoS Comput Biol 3: e42. Accessed 5 March 2012.

78. Guan H, Kiss-Toth E (2008) Advanced technologies for studies on protein interactomes. Adv. Biochem. Eng. Biotechnol. 110: 1–24. Accessed 4 March 2012.

79. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–416. Accessed 1 May 2011.

80. Breitkreutz B-J, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID Interaction Database: 2008 update. Nucleic Acids Res 36: D637–640. Accessed 25 October 2010.

81. Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, et al. (2009) Literature-curated protein interaction datasets. Nat. Methods 6: 39–46. Accessed 7 October 2010.

82. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol. 5: R63. Accessed 14 April 2012.

83. O'Brien KP, Remm M, Sonnhammer ELL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. 33: D476–480. Accessed 14 April 2012.

84. De Lichtenberg U, Jensen LJ, Brunak S, Bork P (2005) Dynamic Complex Formation During the Yeast Cell Cycle. Science 307: 724–727. Accessed 13 April 2012.

85. Mering C von, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein|[ndash]|protein interactions. Nature 417: 399. Accessed 2 March 2012.

86. Albert, Jeong, Barabasi (2000) Error and attack tolerance of complex networks. Nature 406: 378–382. Accessed 8 May 2010.

87. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7: e1001273. Accessed 1 May 2011.

88. De Silva E, Stumpf MPH (2005) Complex Networks and Simple Models in Biology. J. R. Soc. Interface 2: 419–430. Accessed 13 April 2012.

89. Pržulj N (2011) Protein-protein interactions: making sense of networks via graph-theoretic modeling. Bioessays 33: 115–123. Accessed 7 July 2011.

90. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, et al. (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet 41: 1313–1318. Accessed 4 March 2010.

91. De Jager PL, Jia X, Wang J, de Bakker PIW, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat. Genet 41: 776–782. Accessed 19 March 2010.

92. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, et al. (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat. Genet 41: 1228–1233. Accessed 19 March 2010.

93. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet 40: 395–402. Accessed 19 March 2010.

94. Raychaudhuri S (2010) Recent advances in the genetics of rheumatoid arthritis. Curr Opin Rheumatol 22: 109–118. Accessed 4 March 2010.

95. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42: 105–116. Accessed 18 August 2010.

96. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al. (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet 40: 189–197. Accessed 3 August 2010.

97. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40: 584–591. Accessed 28 July 2010.

98. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, et al. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 5: e1000534. Accessed 4 March 2010.

99. Wang K, Li M, Bucan M (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am. J. Hum. Genet 81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17966091. Accessed 3 March 2010.

100. Brunner HG, van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5: 545–551. Accessed 12 March 2010.

101. D'Andrea AD, Grompe M (2003) The Fanconi anaemia/BRCA pathway. Nat Rev Cancer 3: 23–34. Accessed 12 March 2010.

102. Walhout AJM, Reboul J, Shtanko O, Bertin N, Vaglio P, et al. (2002) Integrating Interactome, Phenome, and Transcriptome Mapping Data for the C. elegans Germline. Current Biology 12: 1952–1958. Accessed 12 March 2010.

103. Li L, Zhang K, Lee J, Cordes S, Davis DP, et al. (2009) Discovering cancer genes by integrating network and functional properties. BMC Med Genomics 2: 61. Accessed 30 August 2010.

104. Sengupta U, Ukil S, Dimitrova N, Agrawal S (2009) Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. PLoS ONE 4: e8100. Accessed 30 August 2010.

105. Gentleman R, Huber W (2007) Making the most of high-throughput protein-interaction data. Genome Biol 8: 112. Accessed 30 March 2010.

106. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, et al. (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc. Natl. Acad. Sci. U.S.A 105: 20870–20875. Accessed 25 March 2010.

107. Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686–691. Accessed 10 March 2010.

108. Benita Y, Cao Z, Giallourakis C, Li C, Gardet A, et al. (2010) Gene enrichment profiles reveal T cell development, differentiation and lineage specific transcription factors including ZBTB25 as a novel NF-AT repressor. Blood Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20410506. Accessed 8 May 2010.

109. Ziegler A-G, Nepom GT (2010) Prediction and pathogenesis in type 1 diabetes. Immunity 32: 468–478. Accessed 16 July 2010.

110. Bergholdt R, Størling ZM, Lage K, Karlberg EO, Olason PI, et al. (2007) Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. Genome Biol 8: R253. Accessed 16 July 2010.

111. Wu G, Zhu L, Dent JE, Nardini C (2010) A comprehensive molecular interaction map for rheumatoid arthritis. PLoS ONE 5: e10137. Accessed 16 July 2010.

112. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. Nature 393: 440–442. Accessed 18 May 2010.

113. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20453842. Accessed 18 May 2010.

114. Firestein GS (2003) Evolving concepts of rheumatoid arthritis. Nature 423: 356–361. Accessed 19 March 2010.

115. Abraham C, Cho JH (2009) Inflammatory Bowel Disease. N Engl J Med 361: 2066–2078. Accessed 2 April 2010.

116. Abraham C, Cho J (2009) Interleukin-23/Th17 pathways and inflammatory bowel disease. Inflamm Bowel Dis 15: 1090–1100. Accessed 2 April 2010.

117. Brand S (2009) Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. Gut 58: 1152–1167. Accessed 27 February 2010.

118. Cho JH (2008) The genetics and immunopathogenesis of inflammatory bowel disease. Nat. Rev. Immunol 8: 458–466. Accessed 27 February 2010.

119. Criswell LA (2010) Gene discovery in rheumatoid arthritis highlights the CD40/NF-kappaB signaling pathway in disease pathogenesis. Immunol. Rev 233: 55–61. Accessed 5 April 2010.

120. Takeda K, Clausen BE, Kaisho T, Tsujimura T, Terada N, et al. (1999) Enhanced Th1 Activity and Development of Chronic Enterocolitis in Mice Devoid of Stat3 in Macrophages and Neutrophils. Immunity 10: 39–49. Accessed 4 March 2010.

121. Zhang H, Massey D, Tremelling M, Parkes M (2008) Genetics of inflammatory bowel disease: clues to pathogenesis. Br. Med. Bull 87: 17–30. Accessed 27 February 2010.

122. Lee EG, Boone DL, Chai S, Libby SL, Chien M, et al. (2000) Failure to Regulate TNF-Induced NF-kappa B and Cell Death Responses in A20-Deficient Mice. Science 289: 2350–2354. Accessed 9 April 2010.

123. Munroe ME, Bishop GA (2007) A Costimulatory Function for T Cell CD40. J Immunol 178: 671–682. Accessed 9 April 2010.

124. Bottini N, Vang T, Cucca F, Mustelin T (2006) Role of PTPN22 in type 1 diabetes and other autoimmune diseases. Seminars in Immunology 18: 207–213. Accessed 15 April 2010.

125. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. N. Engl. J. Med 359: 2767–2777. Accessed 18 May 2010.

126. Kano S, Sato K, Morishita Y, Vollstedt S, Kim S, et al. (2008) The contribution of transcription factor IRF1 to the interferon-gamma-interleukin 12 signaling axis and TH1 versus TH-17 differentiation of CD4+ T cells. Nat. Immunol 9: 34–41. Accessed 18 May 2010.

127. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320. Accessed 18 August 2010.

128. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2010) The UCSC Genome Browser database: update 2011. Nucleic Acids Res Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20959295. Accessed 7 December 2010.

129. Morita H, Wu J, Zipes DP (2008) The QT syndromes: long and short. Lancet 372: 750–763. Accessed 18 April 2012.

130. Hubner NC, Bird AW, Cox J, Splettstoesser B, Bandilla P, et al. (2010) Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. J. Cell Biol. 189: 739–754. Accessed 18 April 2012.

131. Olsen JV, Macek B, Lange O, Makarov A, Horning S, et al. (2007) Higher-energy C-trap dissociation for peptide modification analysis. Nat. Methods 4: 709–712. Accessed 18 April 2012.

132. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, et al. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol. Cell Proteomics 8: 2759–2769. Accessed 18 April 2012.

133. Marx SO, Kurokawa J, Reiken S, Motoike H, D'Armiento J, et al. (2002) Requirement of a macromolecular signaling complex for beta adrenergic receptor modulation of the KCNQ1-KCNE1 potassium channel. Science 295: 496–499. Accessed 18 April 2012.

134. Berger SI, Ma'ayan A, Iyengar R (2010) Systems pharmacology of arrhythmias. Sci Signal 3: ra30. Accessed 1 July 2011.

135. Wang Q, Curran ME, Splawski I, Burn TC, Millholland JM, et al. (1996) Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias. Nat. Genet. 12: 17–23. Accessed 18 April 2012.

136. Curran ME, Splawski I, Timothy KW, Vincent GM, Green ED, et al. (1995) A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome. Cell 80: 795–803. Accessed 18 April 2012.

137. Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, et al. (2004) Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell 119: 19–31. Accessed 18 April 2012.

138. Vatta M, Ackerman MJ, Ye B, Makielski JC, Ughanze EE, et al. (2006) Mutant caveolin-3 induces persistent late sodium current and is associated with long-QT syndrome. Circulation 114: 2104–2112. Accessed 18 April 2012.

139. Ueda K, Valdivia C, Medeiros-Domingo A, Tester DJ, Vatta M, et al. (2008) Syntrophin mutation associated with long QT syndrome through activation of the nNOS-SCN5A macromolecular complex. Proc. Natl. Acad. Sci. U.S.A. 105: 9355–9360. Accessed 18 April 2012.

140. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636. Accessed 18 April 2012.

141. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643. Accessed 18 April 2012.

142. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26: 1367–1372. Accessed 18 April 2012.

143. Müller CS, Haupt A, Bildl W, Schindler J, Knaus H-G, et al. (2010) Quantitative proteomics of the Cav2 channel nano-environments in the mammalian brain. Proc. Natl. Acad. Sci. U.S.A. 107: 14950–14957. Accessed 18 April 2012.

144. Rossin EJ, Lundby A, Annette B. S, Christopher N-C, Arne P (2011) Proteomic and genetic dissection of cardiac repolarization protein complexes. Presubmission inquiry sent to Nature .

145. Milan DJ, Kim AM, Winterfield JR, Jones IL, Pfeufer A, et al. (2009) Drug-sensitized zebrafish screen identifies multiple genes, including GINS3, as regulators of myocardial repolarization. Circulation 120: 553–559. Accessed 18 April 2012.

146. Yoshida M, Minamisawa S, Shimura M, Komazaki S, Kume H, et al. (2005) Impaired Ca2+ store functions in skeletal and cardiac muscle cells from sarcalumenin-deficient mice. J. Biol. Chem. 280: 3500–3506. Accessed 18 April 2012.

147. Vasile VC, Edwards WD, Ommen SR, Ackerman MJ (2006) Obstructive hypertrophic cardiomyopathy is associated with reduced expression of vinculin in the intercalated disc. Biochem. Biophys. Res. Commun. 349: 709–715. Accessed 18 April 2012.

148. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322. Accessed 3 August 2011.

149. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet 83: 311–321. Accessed 10 October 2010.

150. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB (2012) The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. PLoS Genet 8: e1002496. Accessed 23 March 2012.

151. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature Available at: http://www.nature.com.ezp-prod1.hul.harvard.edu/nature/journal/vaop/ncurrent/full/nature11011.html. Accessed 5 April 2012.

152. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384. Accessed 10 October 2010.

153. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet 86: 832–838. Accessed 25 October 2010.

154. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89: 82–93. Accessed 23 March 2012.

155. Sakai Y, Shaw CA, Dawson BC, Dugas DV, Al-Mohtaseb Z, et al. (2011) Protein Interactome Reveals Converging Molecular Pathways Among Autism Disorders. Science Translational Medicine 3: 86ra49. Accessed 27 November 2011.

156. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1: S233–240. Accessed 4 October 2010.

157. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24: i223–231. Accessed 1 October 2010.

158. Komurov K, White MA, Ram PT (2010) Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. PLoS Comput. Biol 6. Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20808879. Accessed 18 October 2010.

159. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. 80: 727–739. Accessed 15 February 2012.

160. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. PLoS Comput. Biol. 3: e254. Accessed 15 February 2012.

161. Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and Predicting Haploinsufficiency in the Human Genome. PLoS Genet 6: e1001154. Accessed 15 February 2012.

162. Kazantseva A, Sepp M, Kazantseva J, Sadam H, Pruunsild P, et al. (2009) N-terminally truncated BAF57 isoforms contribute to the diversity of SWI/SNF complexes in neurons. J. Neurochem. 109: 807–818. Accessed 15 February 2012.

163. Kitagawa H, Fujiki R, Yoshimura K, Mezaki Y, Uematsu Y, et al. (2003) The Chromatin-Remodeling Complex WINAC Targets a Nuclear Receptor to Promoters and Is Impaired in Williams Syndrome. Cell 113: 905–917. Accessed 15 February 2012.

164. Yang S-H, Jaffray E, Hay RT, Sharrocks AD (2003) Dynamic interplay of the SUMO and ERK pathways in regulating Elk-1 transcriptional activity. Mol. Cell 12: 63–74. Accessed 15 April 2012.

165. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature Available at: http://www.ncbi.nlm.nih.gov/pubmed/22495309. Accessed 17 April 2012.

166. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, et al. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection. The American Journal of Human Genetics 90: 720–726. Accessed 18 April 2012.

167. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS Genet 7: e1002254. Accessed 18 April 2012.

168. Irvin MR, Shrestha S, Chen Y-DI, Wiener HW, Haritunians T, et al. (2011) Genes linked to energy metabolism and immunoregulatory mechanisms are associated with subcutaneous adipose tissue distribution in HIV-infected men. Pharmacogenetics and Genomics 21: 798–807. Accessed 18 April 2012.

169. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, et al. (2011) Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. 43: 969–976. Accessed 25 April 2012.

170. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, et al. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. 30: 159–164. Accessed 25 April 2012.

171. Dutkowski J, Ideker T (2011) Protein networks as logic functions in development and cancer. PLoS Comput. Biol. 7: e1002180. Accessed 25 April 2012.

172. Sang L, Miller JJ, Corbit KC, Giles RH, Brauer MJ, et al. (2011) Mapping the NPHP-JBTS-MKS protein network reveals ciliopathy disease genes and pathways. Cell 145: 513–528. Accessed 25 April 2012.

173. Amit I, Regev A, Hacohen N (2011) Strategies to discover regulatory circuits of the mammalian immune system. Nat. Rev. Immunol. 11: 873–880. Accessed 25 April 2012.

174. Novershtern N, Regev A, Friedman N (2011) Physical Module Networks: an integrative approach for reconstructing transcription regulation. Bioinformatics 27: i177–185. Accessed 25 April 2012.

175. Bandyopadhyay S, Mehta M, Kuo D, Sung M-K, Chuang R, et al. (2010) Rewiring of genetic networks in response to DNA damage. Science 330: 1385–1389. Accessed 25 April 2012.

176. Ideker T, Krogan NJ (2012) Differential network biology. Mol. Syst. Biol. 8: 565. Accessed 20 April 2012.

177. Fliri AF, Loging WT, Volkmann RA (2010) Cause-effect relationships in medicine: a protein network perspective. Trends Pharmacol Sci Available at: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20810173. Accessed 7 October 2010.

178. Schadt EE, Zhang B, Zhu J (2009) Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. Genetica 136: 259–269. Accessed 27 February 2010.